Universidade de Aveiro Departamento de Matemática,
2018

Catarina
Lopes Vale
Rodrigues Monteiro

Efeito da Gestão Florestal na Produtividade do Eucalipto *Globulus* em Áreas de Minifúndio

Catarina
Lopes Vale
Rodrigues Monteiro

Efeito da Gestão Florestal na Produtividade do Eucalipto *Globulus* em Áreas de Minifúndio

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações - ramo de Estatística e Otimização, realizada sob a orientação científica da Doutora Nélia Maria Marques Silva, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro, e sob a coorientação científica da Doutora Isabel Maria Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

o júri / the jury

presidente / **Prof. Doutor Pedro Filipe Pessoa Macedo**
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais / examiners committee **Prof. Doutora Maria Cristina Souto de Miranda**
Professora Adjunta do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro

Prof. Doutora Nélia Maria Marques da Silva
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

agradecimentos

A realização do presente relatório de estágio contou com apoios e incentivos primordiais, sem os quais não se teria tornado uma realidade. Sentir-me-ei eternamente grata a todas as pessoas que, direta ou indiretamente, contribuíram para a concretização deste sonho.

Em primeiro lugar, quero agradecer aos meus pais, que foram a minha maior inspiração na conclusão desta etapa. São o meu porto seguro, os meus pilares, e a minha maior força, impulsionando-me a ser cada vez melhor a nível pessoal e profissional. Não há palavras suficientes para vos agradecer tudo o que representam para mim. Devo-vos o que sou, e, por isso, estarei eternamente grata.

Aos meus irmãos, pois apesar da tenra idade, tiveram sempre a capacidade extraordinária de perceber quando uma palavra ou um gesto de carinho eram a solução que precisava para os meus complexos problemas. Obrigada por entenderem sempre as razões da minha ausência, e por me ensinarem a olhar o mundo de uma forma mais colorida.

Ao Guilherme, por ser quem é. Por me animar nos momentos de desânimo, pelas palavras e gestos de incentivo e por acreditar incondicionalmente que eu era capaz, mesmo quando nem eu própria acreditava em mim. Obrigada pela paciência, pelo apoio, por todo o amor e por tornares os meus dias mais divertidos. És tudo.

À Mónica, a melhor companheira de estágio que poderia ter tido. Obrigada por todo o companheirismo, amizade, ajuda e cumplicidade. Foste, sem dúvida, uma das pessoas que mais me impulsionou na conclusão desta etapa. As conversas e todo o conhecimento que partilhamos ajudaram-me a alargar os meus horizontes e a crescer enquanto pessoa. Estes anos sem ti não teriam tido o mesmo brilho. Guardarei com carinho todos os momentos.

Às minhas orientadoras, a Doutora Nélia Silva e a Doutora Isabel Pereira, agradeço a disponibilidade, a partilha do saber e as valiosas contribuições para o trabalho. Obrigada por me acompanharem nesta jornada e por estimularem o meu interesse pelo conhecimento e pela Estatística.

À entidade de acolhimento, o instituto de investigação RAIZ, por me terem acolhido durante os longos meses de estágio, pela disponibilidade no esclarecimento de dúvidas e pelo acolhimento ímpar. Em particular, agradeço à Eng.ª Margarida Silva, pela orientação, compreensão e ajuda ao longo dos 8 meses de estágio.

À Filipa, à Rita e à Inês, amigas de sempre, agradeço as longas conversas, os telefonemas e preocupação, assim como o incentivo no desenvolvimento deste trabalho. Às amigas que Aveiro juntou, em particular à Cristiana, à Catarina e à Cláudia, a melhor 'famelga' da vida académica que poderia ter tido.

O meu profundo e sincero agradecimento a todos os que tornaram este projeto possível.

palavras-chave

Amostragem em Inventário Florestal, Gestão florestal, Inferência Estatística, ANOVA não paramétrica a dois fatores, Análise em Componentes Principais, Testes não Paramétricos, Análise de Regressão.

resumo

O presente relatório, elaborado no âmbito do curso de Mestrado em Matemática e Aplicações, no ramo de Estatística e Otimização, da Universidade de Aveiro, pretende reproduzir o estágio curricular realizado no Instituto de Investigação da floresta e papel, RAIZ.

O principal objetivo inerente ao estágio e ao relatório é evidenciar os benefícios económicos da gestão florestal em áreas de minifúndio, mormente na otimização da produtividade nas referidas áreas, cuja espécie predominante é o *Eucalyptus Globulus*, devido à sua importância na produção de pasta e papel.

Para dar resposta ao problema, aplica-se Análise em Componentes Principais, o que permite conjugar as informações de todas as variáveis referentes à produtividade, sendo obtida uma só variável. A partir desta, efetuam-se comparações entre os diferentes grupos de gestão estabelecidos nos dados com base na existência/inexistência de gestão florestal e na entidade responsável pela parcela. As comparações são feitas recorrendo a testes não paramétricos: teste de Kruskal-Wallis, teste de Wilcoxon-Mann-Whitney e ANOVA a dois fatores não paramétrica. Usa-se por fim a análise de regressão para verificar quais os fatores que mais influenciam a produtividade de uma parcela.

keywords Forest Inventory Sampling, Forest Management, Statistical Inference, two way non parametric ANOVA, Principal Components Analisys, Non Parametric Tests, regression analysis.

abstract This report, conducted under the master degree on Mathematics and Applications, in the field of Statistics and Optimization, from Aveiro University, aims to replicate the traineeship held in Instituto de Investigação da floresta e papel, RAIZ.
The main objective inherent to the traineeship and to the report is to highlight the economic benefits of forest management in minifundio areas, mainly in the optimization of productivity in these areas, whose predominant specie is *Eucalyptus Globulus*, due to its importance in the production of pulp and paper.
To answer the problem, it is applied Analysis in Principal Components, which allows to combine the information of all the variables related to the productivity, from which is obtained a single variable. From this, comparisons are made between the different management groups established in the data based on the existence / nonexistence of forest management and on the entity responsible for the area. Comparisons are made using non-parametric tests: Kruskal-Wallis test, Wilcoxon-Mann-Whitney test and non-parametric two-way ANOVA. Finally, the regression analysis is used to verify the factors that most influence the productivity from the areas.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

No cerne da atualidade, o ambiente, a sustentabilidade e a gestão florestal ganham uma importância reconhecida e reconhecível de âmbito alargado, assumindo um papel preponderante em inúmeras unidades da economia, mormente, no setor florestal. Este evidencia-se como sendo um dos setores económicos de maior importância para Portugal. A sua atividade centra-se e promove três grandes pilares da sustentabilidade: económico, social e ambiental. A nível económico, salienta-se a indústria da pasta, papel e cartão, a qual possui relevante interesse estratégico para Portugal, nomeadamente na economia portuguesa, com especial realce para o mercado de exportações, sendo um dos maiores exportadores de valor acrescentado nacional.

Considerando os crescentes desafios que Portugal enfrenta, inúmeras entidades de destaque consideram que o papel produzido através do setor florestal pode ser incrementado quantitativa e qualitativamente, assumindo uma posição cada vez mais relevante na economia nacional, contribuindo fortemente para o crescimento e desenvolvimento económico do País. Sendo Portugal um país pautado por condições de solo e clima de excelência, que permitem no quadro europeu, um desenvolvimento florestal competitivo, a que acresce a disponibilidade de área territorial, é primordial o desenvolvimento de técnicas de gestão florestal que maximizem a produtividade. Com efeito, tal permite evidenciar o potencial de crescimento deste setor nos três pilares de sustentabilidade supracitados. Alguns empresários e gestores de microempresas questionam frequentemente a importância e finalidade da gestão florestal. Há *inclusive* quem desconheça a possibilidade de gerir as suas parcelas por meio de tecnologias de última geração, e consultores especializados em planeamento, operações e produtividade.

No cerne da economia da pasta e do papel, destaca-se um dos três maiores exportadores de Portugal e um dos principais criadores de riqueza para o País (representando 1% do PIB nacional), [1]: a *Navigator Company*. Com efeito, o grupo é responsável pela gestão em cerca de 3,2% da área florestal Portuguesa, e 14,4% da floresta nacional de eucalipto, [2].

No âmbito da realização do estágio curricular conferente ao grau de Mestre em Matemática e Aplicações, no ramo de Estatística e Otimização, a empresa por mim escolhida foi o Centro de Investigação da Floresta e Papel, RAIZ. Os projetos desenvolvidos por este Centro de Investigação são maioritariamente da responsabilidade da multinacional *Navigator Company*. A minha opção na escolha desta empresa para o desenvolvimento do estágio curricular prendeu-se com o reconhecido valor e posicionamento a nível nacional em termos florestais. Nesse contexto, destaca-se a sua preocupação, em termos económicos, e na prossecução da salvaguarda da

floresta, em questões relacionadas com a sua sustentabilidade. Com efeito, a investigação desenvolvida pelo Instituto de Investigação florestal da pasta e papel RAIZ, pauta-se pela preocupação ambiental crescente, aliada à sua sustentabilidade e crescente preocupação na prevenção de incêndios florestais. Assim, os princípios inerentes aos projetos desenvolvidos por este Instituto de Investigação coadunam-se com alguns dos princípios por mim defendidos no que concerne à defesa, proteção e manutenção da floresta. Tal, motivou e incentivou a minha escolha. Paralelamente, a importância crescente que pauta a temática relativa à gestão florestal que se encontra cada vez mais no cerne da atualidade, incitaram a escolha da temática inerente à presente investigação.

A necessidade crescente de implementação de gestão florestal por forma a garantir a sustentabilidade ambiental, a maximização da produtividade, a prevenção e combate de incêndios e a garantia de qualidade paisagística motivaram e estimularam a escolha do tema da tese. Tomar consciência de como a Matemática, especialmente a Estatística, influencia e abrange positivamente estas questões da gestão controlada utilizando metodologias científicas abrangentes, possibilita o incremento substancial da produtividade florestal apresentando-se como tema central da presente investigação.

Assim, o presente estudo objetiva evidenciar os benefícios económicos da gestão florestal em áreas de minifúndio, mormente na otimização da produtividade das referidas áreas. As plantações predominantes nessas áreas são o *Eucalyptus Globulus*, na medida em que esta espécie se apresenta como matéria prima primordial na produção de pasta e papel, o que incita a sua utilização pelos principais produtores de pasta e papel, em particular, a *Navigator Company*.

O desenvolvimento da presente investigação tem por base a aplicação de várias técnicas estatísticas, implementadas com recurso ao *software* R (versão 3.4.2).

Para dar resposta ao problema proposto pela entidade de acolhimento, usaram-se conjuntos de dados, para os quais a implementação de metodologias estatísticas se mostra profícua na obtenção de uma solução adequada para o problema em questão. Os conjuntos de dados fornecidos são constituídos por informações relativas ao inventário florestal de parcelas pertencentes a três entidades, designadas, de uma forma aleatória, por A, B e C, que, por uma questão de confidencialidade, não são identificadas na presente investigação.

No que concerne à estrutura da presente dissertação, primeiramente apresenta-se uma introdução da empresa escolhida para o cumprimento do estágio curricular - o instituto de Investigação da Floresta e Papel, RAIZ - analisando-se a importância da realização de inventário florestal e respetivo lançamento. Nesse âmbito, foi recolhida informação que se encontrava dispersa e foi desenvolvida uma síntese do procedimento usado para o lançamento do inventário florestal, que deu origem ao conjunto de dados fornecido, por forma a, neste contexto, propor sugestões de melhoria para a organização de acolhimento.

No capítulo subsequente são apresentados os conceitos teóricos inerentes às metodologias estatísticas implementadas no capítulo 5 para dar resposta ao problema proposto pela organização de acolhimento. Assim, a abordagem dos capítulos teóricos mostrou-se profícua na medida em que permitiu tomar consciência das metodologias existentes e, consequentemente, escolher as mais adequadas para a resolução do problema em estudo, permitindo a sua análise de forma detalhada.

No capítulo 4 é explorado o problema proposto pela entidade de acolhimento, facilitando o

seu entendimento. São ainda analisadas as variáveis do conjunto de dados em estudo, assim como características associadas às áreas de estudo, para assim proceder à análise detalhada das condições inerentes ao problema.

No seguimento dos capítulos supracitados, foram desenvolvidas análises exploratórias ao conjunto de dados, o que permitirá dar resposta ao problema em questão. Assim, primeiramente foi implementada uma análise preliminar dos dados, o que possibilita coadunar as análises subsequentes ao objetivo explanado. Consequentemente, a análise preliminar dos dados desenvolvida, permitiu constatar a existência de variáveis indicativas da produtividade da parcela, sobre as quais é aplicada Análise em Componentes Principais (ACP), cujo uso se mostrou frutífero na medida em que essas variáveis apresentam valores de correlação significativos. A sua aplicação viabiliza a obtenção de uma só variável - produtividade - sobre a qual as análises subsequentes incidem. Na comparação da produtividade entre parcelas onde existem evidencias de gestão florestal com parcelas onde não se verifica a existência de gestão florestal, foram formados cinco grupos de parcelas, os quais apresentam metodologias de gestão florestal distintas. A sua formação conjugou duas informações preponderantes, a saber: existência ou inexistência de gestão florestal, e a entidade responsável pela parcela.

Assim, objetivando dar resposta ao problema em questão, é primordial comparar a produtividade entre as parcelas pertencentes aos cinco grupos formados. Para tal ir-se-á recorrer a testes não paramétricos: o teste de Kruskal-Wallis, teste de Wilcoxon-Mann-Whitney e ANOVA a dois fatores não paramétrica. A implementação e verificação da última metodologia supracitada será por mim desenvolvida no *software* R, por forma a que seja obtida uma função capaz de ser generalizada para outros conjuntos de dados. A inadequação da aplicação de testes paramétricos ao conjunto de dados em estudo, justifica-se com base na falta de Normalidade da variável produtividade em alguns grupos em estudo. Por fim, a aplicação de metodologias de regressão adequadas permitirá intuir quanto à influência de um conjunto de variáveis (de índole quantitativa e qualitativa) na produtividade das parcelas.

A título conclusivo enfatiza-se a importância e valorização da gestão florestal na preservação, manutenção e desenvolvimento do setor florestal, essenciais no incremento da produtividade das parcelas e na prevenção de incêndios florestais. Ainda a relevância na maximização dos benefícios económicos relacionados com a indústria do papel, pasta e cartão, sendo preponderante a adequação das técnicas aplicadas, coadunando-se com a tipologia de solo e clima existentes nas áreas exploradas.

Capítulo 2

O estágio curricular

"O estágio enobrece, traz conhecimento e dignifica o ser humano"

Clarice Lispector

No presente capítulo são brevemente explorados diversos conceitos inerentes ao estágio curricular realizado. Assim, primeiramente efetua-se uma breve apresentação da entidade de acolhimento. Seguidamente, tendo em conta que a presente investigação utiliza dados relativos ao inventário florestal, após uma breve descrição dessa temática, é apresentado o procedimento seguido pela entidade de acolhimento no lançamento do inventário florestal que possibilitou a obtenção dos dados em estudo.

2.1 A organização de acolhimento - RAIZ, Instituto de Investigação da floresta e do papel

O RAIZ - Instituto de Investigação da Floresta e do Papel, fundado em 1996 pela *The Navigator Company*, constitui um organismo privado e sem fins lucrativos, [1].

Figura 2.1: A entidade de acolhimento - instituto de investigação da floresta e papel, RAIZ.

A colaboração do RAIZ com vários polos do ensino superior no País - Universidades de Aveiro, Coimbra, Porto, Minho, Trás-os-Montes e Alto Douro, Beira Interior, Instituto Superior de Agronomia; ou, no estrangeiro (Europa, América e Austrália), instituições como o *Cooperative Research Center for Sustainable Production Forestry* - motivou o reconhecimento de objetivos comuns à indústria e ao mundo académico, incrementando a existência de um trabalho conjunto para a realização dos fins a que se propõe, [17, 2]. É ainda membro do mais importante projeto na área florestal – consórcio *Genolyptus*: Rede Brasileira de Pesquisa do Genoma de *Eucalyptus*, pautado pelo reconhecimento mundial, [71].

2.1.1 Missão

A Missão do RAIZ - Instituto de Investigação da Floresta e Papel é contribuir para a competitividade da Fileira do Eucalipto. Objetiva dar resposta à necessidade de identificar, a cada momento, segmentos prioritariamente elegíveis para a atividade de investigação aplicada, visando otimizar numa ótica de custo/benefício, as vantagens competitivas da fileira silvo-industrial nacional garantindo a sua sustentabilidade. Com efeito, o foco do RAIZ vai mais além do incremento da produtividade florestal, centrando-se também no aumento da qualidade da fibra produzida (por forma a reduzir custos de produtividade e, concomitantemente, aumentar a qualidade do papel - a partir do melhoramento genético do eucalipto), implementando uma gestão florestal sustentada adaptada a cada região, [63].

A investigação aplicada nos domínios da produção florestal e da tecnologia industrial da pasta e papel, surge inerente à missão do RAIZ, sendo esta desenvolvida numa perspetiva de transformar conhecimentos em tecnologia, [2].

2.1.2 Visão

Tem como objetivo a obtenção de reconhecimento, a nível mundial, como um centro de investigação de referência, promotor do desenvolvimento sustentável e da bioeconomia baseada na floresta do eucalipto, constituindo o maior repositório mundial de conhecimento científico e tecnológico da espécie *Eucalyptus Globulus*, sendo a mais predominante em Portugal, [1].

2.1.3 Atribuições e competências

A atividade do RAIZ possui três âmbitos de intervenção principais: **Investigação Aplicada**, **Consultoria** e **Formação**.

Relativamente ao âmbito da **Investigação Aplicada**, salientam-se duas áreas de intervenção: tecnológica e florestal. O seu desenvolvimento deriva de uma estreita colaboração com a indústria e em função de objetivos definidos concretizados na forma de projetos, cada um dos quais conduzido por um gestor responsável pela coordenação de toda a atividade. As parcerias estabelecidas pelo RAIZ são responsáveis pelo seu posicionamento numa situação privilegiada para desenvolver, com sucesso, a sua atividade de investigação aplicada. Assim, primeiramente os projetos são submetidos à aprovação da entidade financiadora, sendo também analisados numa perspetiva científica pelo Conselho Científico do RAIZ. A atividade de **Consultoria** é realizada para diversas entidades das áreas florestal e tecnológica, não sendo restrita para as

empresas sócias.

Objetivando a criação de condições que fomentem a disponibilização de quadros altamente especializados e com competência reconhecida a nível nacional e internacional em todas as áreas relevantes das fileiras florestal e tecnológica do eucalipto, a atividade do RAIZ possui uma forte componente na **Formação** dos seus trabalhadores.

Salienta-se ainda a vasta experiência no estudo de pragas do eucalipto, no melhoramento genético do eucalipto e biotecnologia, o que confere ao RAIZ a capacidade técnica para estruturar e executar os projetos aos quais se propõe, [2, 63].

2.1.4 Prémios e reconhecimentos

O trabalho desenvolvido permitiu à empresa ver o seu êxito reconhecido, tendo sido distinguida pelos seus 20 anos de trabalho de investigação do eucalipto com um galardão atribuído pela ordem dos engenheiros, bem como o seu reconhecimento como entidade do Sistema Científico e Tecnológico Nacional. Salienta-se ainda a contribuição do instituto de investigação na melhoria das performances da *Navigator Company*, nomeadamente assegurando "elevados níveis de eficiência de produção e qualidade da pasta" e ajudando a "diversificar o portefólio de produtos da empresa por implementação do conceito de bio refinaria", [17].

2.1.5 Recursos físicos e humanos

A sede do RAIZ está inserida na Quinta de São Francisco, a 9 km de Aveiro, onde o instituto de investigação desenvolve a sua atividade de investigação maioritária. Dispõe de edifícios administrativos, laboratórios de bancada, instalações piloto, serviços centrais da sede, serviços de documentação, instalações para formação, sociais e de acolhimento, o que compreende uma área coberta de 2600 m^2, aproximadamente. Paralelamente, possui ainda viveiros e parques de hibridação e um laboratório de biotecnologia, os quais se situam na Herdade de Espirra, em Pegões, [2, 63].

No âmbito dos recursos humanos reitera-se que, atualmente, a equipa do RAIZ é constituída por 76 colaboradores, dos quais 46 são permanentes e 30 são bolseiros. No que remete aos 76 colaboradores, 73 são responsáveis pelas atividades de Investigação e Desenvolvimento, enquanto que os demais colaboradores se encontram encarregues da parte administrativa.

Relativamente aos 46 colaboradores permanentes, 21 constituem o quadro do RAIZ, enquanto que os restantes 25 fazem parte do quadro do grupo *The Navigator Company*.

2.1.6 O grupo *The Navigator Company*

O grupo *The Navigator Company* detém uma participação de 94% no capital associativo do RAIZ, dedicando-se ao fabrico e comercialização de papel em Portugal de forma totalmente autónoma, nomeadamente no que diz respeito à madeira, pasta e ao papel, [14, 1].

A sua atividade permitiu-lhe ser reconhecido como líder mundial no segmento *premium* de papeis de escritório com a marca *Navigator*, bem como produtor mundial de papeis finos de impressão e escrita não revestidos; e ainda de pasta branqueada de eucalipto, [71].

Assim, sendo um dos três maiores exportadores de Portugal e um dos principais criadores de riqueza para o país (representando 1% do PIB nacional), o grupo *The Navigator Company* exporta quase a totalidade da produção para 118 países dos cinco continentes. O grupo gere 3.2% da área florestal portuguesa e 14.4% da floresta nacional de eucalipto, [76].

2.2 Inventário florestal

Desde tempos primórdios que o modo como o homem interage com a floresta tem sofrido uma panóplia de modificações, [30]. Se, por um lado, as antigas sociedades humanas utilizavam de forma descomedida os recursos florestais, imaginando-a como sendo um recurso inesgotável, desde há alguns séculos que, tendo em conta a desflorestação generalizada e a iminente falta de recursos naturais, tem sido desenvolvida uma preocupação crescente com o meio ambiente e, em particular, com a sustentabilidade florestal, [32, 60, 53, 24].

Atualmente, os produtos e serviços que a sociedade espera da floresta são cada vez mais diversificados. Desta forma, a sociedade cada vez mais exigente obriga ao desenvolvimento de uma gestão florestal pautada pela preocupação crescente não só em manter a produção de lenho numa base sustentada, mas também em garantir a estabilidade do ecossistema florestal e satisfazer as exigências de uma sociedade em permanente evolução, [53, 7]. A intervenção exercida sob a floresta para que dela sejam obtidos os mais diversos produtos e serviços - produtos lenhosos e não lenhosos, recreio e lazer, atividade cinegética, entre outros - tem de ser compatível com a manutenção da sua biodiversidade, produtividade, capacidade de regeneração e vitalidade, de modo a garantir que as gerações vindouras venham a ter igual oportunidade de a utilizar, [54, 74]. Com efeito, estabelece-se uma relação entre o homem e a floresta que implica a gestão florestal sustentada através da qual se garante, concomitantemente, a obtenção dos recursos naturais necessários à sobrevivência do ser humano e à sustentabilidade florestal. Para tal é primordial que os gestores florestais desenvolvam programas de inventariação e monitorização de recursos florestais para que detenham informação fiável na qual se possam basear nas suas opções de gestão e, também, de dados que lhes permitam avaliar, no futuro, as consequências das decisões tomadas, [53]. Assim, o inventário de recursos florestais implica a caracterização de uma determinada área florestal, [7, 24, 67].

2.2.1 Elementos a obter num inventário florestal

Podem realizar-se inventários florestais de diversas escalas e com diversos objetivos, reportando-se também a diferentes tipos de utilizadores, desde proprietários privados que estão preocupados com a gestão do seu povoamento ou com a venda dos seus produtos, até aos administradores públicos e políticos que pretendam uma caracterização do sector florestal que os ajude a definir medidas de política florestal. Pode assim definir-se inventário florestal como o conjunto de procedimentos que permitem caracterizar uma determinada área florestal, tendo em vista um determinado objetivo, [61, 53, 29]. A caracterização da área florestal pode, assim, implicar a obtenção de diferentes elementos. Os elementos que são geralmente tidos em conta para a caracterização de uma área florestal são:

- **Descrição da área florestal e avaliação de áreas:** definição da tipologia dos povoamentos florestais, eventual produção de cartografia, descrições topográfica e hidrográfica,

regime de propriedade, acessibilidade, etc.

- **Caracterização dos povoamentos florestais:** inclui informação dendrométrica e não dendrométrica; a informação dendrométrica implica a medição das árvores – geralmente em parcelas de amostragem - e os cálculos necessários para a estimação das variáveis dendrométricas que caracterizam os povoamentos; a informação não dendrométrica implica a recolha de outra informação importante para a caracterização do povoamento – sub-bosque, regeneração natural, estado sanitário, etc – a qual não implica a medição de árvores.

- **Caracterização dos matos:** inclui a caracterização das zonas de matos, a qual inclui geralmente a recolha da informação não dendrométrica;

- **Avaliação de indicadores de gestão florestal sustentável:** avaliação de um conjunto pré-selecionado de indicadores de gestão florestal sustentável (IGFS) -valor recreativo, a vida silvestre, a diversidade vegetal arbustiva em sub-coberto, a presença de espécies protegidas, o armazenamento de carbono, o perigo de incêndio, a desfolhação, deficiências nutricionais, etc.

- **Avaliação de acréscimos:** estimação dos acréscimos em volume nos últimos anos (o número de anos depende da periodicidade dos inventários).

- **Determinação de perdas:** inclui a estimação da quantidade de madeira que é cortada ou destruída por pragas e doenças nos últimos anos.

O objetivo inerente ao inventário tem influência determinante sobre o relevo que se dá a cada um dos elementos a recolher no inventário. A definição clara dos objetivos que se pretende alcançar na realização de um inventário é, assim, essencial para o seu correto planeamento. Com efeito, um inventário florestal é uma tarefa complexa cujo planeamento implica um conjunto de decisões extremamente importantes, não só ao nível da qualidade do resultado final, mas também em termos de exatidão, tempo e custos envolvidos. O sucesso de um inventário florestal depende indubitavelmente de um planeamento cuidado dos trabalhos a efetuar, [74, 24, 66].

2.3 Lançamento do inventário florestal – RAÍZ

A obtenção dos dados em estudo resultou do lançamento de um inventário florestal, cujo objetivo inerente seria determinar de que forma a gestão florestal implementada em determinada região permite otimizar a sua produtividade. Assim, com base na referida definição clara dos objetivos intrínsecos ao inventário florestal, foram desenvolvidos, pelo RAIZ, um conjunto de procedimentos. Por forma a facilitar a utilização e compreensão dos procedimentos mencionados, foi desenvolvida, ao longo do tempo decorrente do estágio curricular, uma recolha de informação dispersa no setor responsável por esta tarefa. Assim, seguidamente apresenta-se a sistematização resultante desses passos a seguir aquando do lançamento do inventário florestal.

1. **Estratificação das áreas de estudo:**

O primeiro passo no lançamento do inventário florestal consiste na estratificação das áreas de estudo, recorrendo-se, para tal, a critérios que permitam agrupar as parcelas das áreas de estudo em estratos. Este procedimento permite a obtenção de estratos homogéneos no seu interior relativamente a características inerentes às parcelas, e heterogéneos entre si. Desta forma, assegura-se a representatividade dos estratos (e, por conseguinte, das parcelas) selecionados. A estratificação das áreas de estudo é feita, essencialmente, com base nas categorias do uso do solo.

2. **Definição do número de parcelas a lançar:**

   No passo que antecede o lançamento das parcelas, é necessário determinar o número de parcelas adequado para usar no processo de amostragem, garantindo desta forma a representatividade da população em estudo, que, neste caso, designa as uma região do centro de Portugal. O referido procedimento foi implementado por investigadores do RAIZ, tendo sido auxiliado pelo Excel (Anexo A). A determinação do número de parcelas a amostrar, n, recorre aos seguintes parâmetros do modelo:

   (a) Área do estrato: área total do estrato que se pretende amostrar;

   (b) Dimensão da parcela: definição da dimensão que se pretende usar em cada parcela de cada estrato, tendo sido estabelecido, no presente inventário, que cada parcela teria, aproximadamente, 400 m^2.

   (c) Produtividade média: volume médio de produtividade, em m^3/ha.

   (d) Desvio padrão: definido em função do volume médio de produtividade, em cada região de estudo.

   (e) Precisão: o presente estudo utiliza três precisões: 0.1, 0.15 e 0.2, sendo os cálculos implementados para as diferentes precisões.

   (f) Erro admitido (E): para esta variável admite-se um erro de 10%, aproximadamente. O valor do erro depende do valor da precisão, e é calculado com base na seguinte fórmula:

   $$E = precisao \times Volume_{medio} \tag{2.1}$$

   Com efeito, este valor diz respeito ao erro admitido, cujo valor corresponde a mais ou menos o valor da precisão do valor médio do volume.

   (g) O custo (estimado) gasto em cada parcela foi de 20 € (valor fixo).

   (h) Para calcular o número máximo possível de unidades amostrais (parcelas) para cada estrato i, a fórmula usada é a seguinte:

   $$N_i = \frac{Area\,da\,regiao\,i}{Dimensao\,da\,parcela\,i} \tag{2.2}$$

   (i) Sendo N o valor que designa o tamanho da população, ou seja, o máximo número de unidades amostrais possíveis, em todos os estratos, o modelo que permite calcular o valor de N vem:

   $$N = \frac{Area\,total\,das\,regioes}{dimensao\,total\,das\,parcelas} = \sum_{i=1}^{L} N_i \tag{2.3}$$

(j) Após calcular o valor de N, é necessário efetuar o cálculo do número de parcelas a lançar para o total de estratos (n). O cálculo foi efetuado para cada valor de precisão enunciado na alínea e). Para tal, a fórmula usada é:

$$n = \left(\frac{t_{st}}{E}\right)^2 \left[\sum_{i=1}^{L} W_i \times S_i \times \sqrt{C_i}\right] \times \left[\sum_{i=1}^{L} W_i \times S_i / \sqrt{C_i}\right] \tag{2.4}$$

(k) Para cada estrato é necessário também calcular o número de parcelas a lançar, n_i. Assim, para cada valor de precisão é calculado o número de parcelas lançar em cada estrato através da seguinte fórmula:

$$n_i = n \times \frac{W_i \times S_i / \sqrt{C_i}}{\sum_{i=1}^{L} W_i \times S_i / \sqrt{C_i}} \tag{2.5}$$

Onde, as variáveis usadas nas equações anteriores correspondem a:

- i=1,. . . ,L os estratos;
- ***L:*** o número total de estratos determinados;
- t_{st}***:*** *t-Student* para um intervalo de confiança de 95%, com 2 graus de liberdade;
- ***E:*** erro admitido (± precisão do volume médio);
- S_i***:*** desvio padrão do estrato i;
- n_i***:*** número de unidades amostrais a medir no estrato i que é alocado de forma proporcional a $W_i \times S_i / \sqrt{C_i}$.
  Se $n_i < 3$, fazer $n_i = 3$.
  Tem-se ainda que $W_i = \frac{N_i}{N}$.
- ***n:*** é o número total de unidades amostrais (parcelas) a serem medidas no número total de estratos.
- N_i***:*** número máximo possível de unidades amostrais (parcelas) para cada estrato i, calculado a partir do quociente entre a área do estrato i, e pela dimensão média da parcela i .
- ***N:*** tamanho da população ou máximo número de unidades amostrais possíveis do número total de estratos;
- C_i***:*** custo de selecionar e medir um polígono do estrato i.

3. **Lançamento das parcelas:**

   O lançamento das parcelas é feito com recurso ao *software* QGIS, com base na ferramenta *'random points inside polygon variable'*. Esta metodologia parte da estratificação feita, e, através dos polígonos formados pelo *software*, lança pontos dentro dos polígonos, de forma aleatória. A partir das coordenadas lançadas pelos referidos pontos, são estabelecidas as parcelas a inventariar. Estas parcelas possuem forma (aproximadamente) circular, onde o ponto lançado aleatoriamente corresponde ao centro da parcela, e o raio é aproximadamente de 12m, o que permite perfazer uma área de parcela de, aproximadamente, 400 m^2, tal como pretendido.
   Em certos casos, o ponto aleatório resultante da aplicação do *software* é lançado sobre locais que, geograficamente, não tornam o desenvolvimento do inventário exequível ou que adulteram os resultados, a saber:

(a) Lançamento do ponto aleatório em cursos de água (visto que essas áreas favorecem a produtividade dada a afluência de recursos hídricos);

(b) Parcela muito próxima de estradas ou zonas onde predominam aberturas da floresta que favorecem a entrada de luz solar nestes locais (na medida a que a entrada de luz solar favorece a produtividade dessas áreas);

(c) Locais de difícil acesso, mormente zonas montanhosas ou depressões;

Por forma a tornar os resultados obtidos plausíveis e realísticos, tanto quanto possível, eliminando, por conseguinte, a existência de *outliers* e de resultados que não traduzem devidamente a realidade, o ponto aleatório, e, por conseguinte, a área a inventariar, é deslocada para zonas fidedignas. As metodologias supracitadas são implementadas em trabalho de campo, pelo grupo que procede à recolha de dados, ficando as alterações efetuadas devidamente registadas.

4. **Medição Biométrica das parcelas:**

   Após o lançamento das parcelas a inventariar, é necessário efetuar as medições em campo, por um grupo especializado de trabalhadores. A descrição das variáveis encontra-se na secção 4.3 passível de ser consultada, sendo que as variáveis resultantes da medição biométrica das parcelas possuem essa indicação.

5. **Processamento:**
   Com base nos dados recolhidos, isto é, a partir das variáveis medidas em trabalho de campo, são efetuados cálculos que permitem inferir quanto a medidas de qualidade da parcela e outras variáveis de relevo. As fórmulas para os respetivos cálculos encontram-se tabeladas e são comummente usadas na área da biometria. O instituto de investigação da florestal e papel efetuou a sua implementação no *software* R, permitindo o seu uso generalizado, e simplificando este procedimento. As variáveis cujos valores são obtidos pela metodologia supracitada encontram-se indicadas na tabela 4.1 da secção 4.3.

Assim, em suma, o procedimento seguido pelos investigadores do RAIZ no lançamento do inventário florestal e recolha dos respetivos dados é o seguinte:

- **Estratificação** das áreas que são objeto de estudo, permitindo a obtenção de estratos cujas parcelas são semelhantes entre si relativamente a uma ou mais características. Os estratos devem garantir a representatividade das áreas em estudo.
- **Seleção do estrato** para efetuar o lançamento aleatório das parcelas.
- **Lançamento aleatório das parcelas em cada estrato** - utilizando para tal uma Função/algoritmo do *software* QGIS - *'random points inside polygon variable'*. Cada ponto (coordenada geográfica) é lançado dentro de cada polígono (estrato) do GGIS, de forma aleatória.
- **Lançamento de N parcelas**, de forma aleatória, sobre cada estrato: são lançados tantos pontos quanto o número de parcelas calculadas para cada estrato (N). As parcelas são construídas em torno de cada ponto (coordenada obtida pela função do QGIS), assumindo uma forma aproximadamente circular, cujo raio é tal que cada parcela possua, aproximadamente, 400 m^2 de área.

- O procedimento explanado é repetido para cada estrato estabelecido.
- **Correção de pontos aleatórios** lançados pelo *software* (trabalho de campo): a existência de valores considerados enganosos nos estudos a realizar ou que não sejam de fácil acesso, instigam o deslocamento das parcelas para locais próximos dos lançados pelo *software* e cujos resultados sejam fidedignos.
- **Medição biométrica** dos dados em campo, e **cálculo dos restantes valores** necessários ao estudo das parcelas, utilizando as respetivas fórmulas programadas no *software* R.

Capítulo 3

Revisão da Metodologia Estatística

"O conhecimento matemático acrescenta vigor à mente, liberta-a de preconceitos, da superstição e da credulidade."

John Arbuthnot

No presente capítulo faz-se uma breve revisão das noções, métodos e resultados da análise estatística que serviram de suporte às técnicas utilizadas no seguimento da investigação desenvolvida. Desta forma, a exposição destina-se a assegurar a compreensão dos métodos e dos resultados em discussão.

3.1 Análise em Componentes Principais (ACP)

O estudo e interpretação da estrutura de variâncias-covariâncias de um determinado fenómeno, medido por múltiplas variáveis de modo a revelar relações entre variáveis, entre sujeitos, e entre sujeitos e variáveis, é o objetivo primordial da ACP. A análise em componentes principais (ACP) é uma técnica de análise exploratória multivariada que, a partir de um conjunto de variáveis correlacionadas entre si, permite obter um novo conjunto de dados preferencialmente de dimensão inferior ao conjunto de dados original, e cujas variáveis são não correlacionadas entre si, [21, 48]. Assim, o seu objetivo é encontrar uma transformação ortogonal das variáveis originais, que defina um novo conjunto de variáveis, não correlacionadas entre si, e que permitam explicar a maior quantidade de variabilidade dos dados possível. Às variáveis ortogonais resultantes da aplicação de ACP, designam-se por Componentes Principais (CP), as quais resumem a informação disponível nas variáveis originais, [31].

Descrita desta forma, a ACP é comummente considerada como uma metodologia de redução da dimensionalidade dos dados, pelo que uma das principais utilizações da ACP é o resumo da informação de várias variáveis correlacionadas (e, portanto, de alguma forma redundantes) em uma ou mais combinações lineares independentes, as componentes principais, que representam a maior parte da informação presente nas variáveis originais. Note-se que caso as variáveis originais não fossem correlacionadas, a sua aplicação não acrescentaria conhecimentos adicionais, [50, 52].

Finalmente, destaca-se uma questão de terminologia inglesa usada nos programas de cálculo estatístico: os coeficientes lineares que definem as CPs são designados por *loadings* e os coeficientes de cada indivíduo numa CP são designados por *scores*, [52].

3.1.1 Estimação das componentes principais

Ao nível algébrico, as componentes principais são combinações lineares de p variáveis originais correlacionadas, sendo que a análise em componentes principais (ACP) permite formar p combinações lineares independentes, [52, 38].

Seja $X = \left(X_1, X_2, \ldots, X_p\right)^T$ o vetor das variáveis originais que foram observadas na amostra de dimensão n, seja Σ a sua matriz de covariâncias (com $det\,(\Sigma) > 0$), e sejam $\lambda_1 > \cdots > \lambda_p$ os valores próprios de Σ, ordenados por ordem decrescente. As componentes principais são combinações lineares dessas p variáveis definidas por:

$$CP_j = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{pj}X_p = e_j^T X\,, \tag{3.1}$$

onde $j = 1, 2, \ldots, p$ e $e_j^T = (e_{1j}, e_{2j}, \ldots, e_{pj})$ são vetores de constantes de norma unitária, ou seja, tais que $e_j^T e_j = 1$, para todo o j e que $Corr\left(CP_i, CP_j\right) = 0$, para qualquer par (i, j), com $i \neq j$, $i, j = 1, 2, \ldots, p$.

As componentes principais (CP) verificam as seguintes propriedades, [10, 77]:

1. A variância da j-ésima CP é igual ao j-ésimo valor próprio da matriz de covariâncias Σ, isto é,

$$Var(CP_j) = e_j^T \Sigma e_j = \lambda_j, \quad 1 \leq j \leq p. \tag{3.2}$$

2. A soma das variâncias da CPs é igual à soma das variâncias das variáveis originais e igual à soma dos valores próprios de Σ, isto é,

$$\sum_{j=1}^{p} Var\left(CP_j\right) = \sum_{j=1}^{p} Var\left(X_j\right) = \sum_{j=1}^{p} \lambda_j = tr(\Sigma). \tag{3.3}$$

3. O coeficiente de correlação (de Pearson) entre a variável original X_k e a j-ésima CP é dado por:

$$\rho_{k,j} = e_{kj} \frac{\sqrt{\lambda_j}}{\sqrt{Var\left(X_k\right)}} \quad k, j = 1, 2, \ldots, p \quad j \neq k. \tag{3.4}$$

4. $e_j^T e_j = 1$, $j = 1, 2, \ldots, p$, isto é, o vetor e_j^T tem norma unitária.

A decisão sobre o número de componentes principais a considerar depende da percentagem de variância explicada pelas primeiras k componentes principais, sendo, pois, uma questão pautada de alguma subjetividade, [21, 78, 38]. Existem, no entanto, critérios para selecionar o número de componentes a reter:

1. Incluir o número mínimo de componentes que expliquem 85% da variância total, [48].
2. Reter as componentes cujas variâncias são maiores do que um, [21].
3. Decidir com base na representação gráfica, por ordem decrescente, da percentagem de variação total explicada por cada componente (gráfico do cotovelo), [52].

Em termos amostrais, as CPs são estimadas usando as médias, variâncias e correlações amostrais (corrigidas), consoante o caso, [10, 77].

3.2 Testes paramétricos

A comparação de parâmetros populacionais (média, variância, mediana, entre outros) a partir de amostras aleatórias é uma das necessidades mais frequentes em análise estatística. Este tipo de inferência estatística é particularmente útil para testar a significância de fatores que são capazes de influenciar a resposta da variável dependente, na qual se pretende testar se o fator em estudo tem ou não um efeito significativo, [52, 57].

Existem, basicamente, duas metodologias para efetuar este tipo de testes: os testes paramétricos e testes não paramétricos. Enquanto que os primeiros exigem que a forma da distribuição da população seja conhecida (a Normal é a mais utilizada), os segundos não exigem à partida o conhecimento da distribuição amostral (o que não implica, porém, que estes testes não possuam outras condições de aplicação). Não obstante e de uma forma geral, a potência dos testes paramétricos é superior à dos testes não paramétricos, ou seja, a probabilidade de rejeitar, corretamente, a hipótese nula é maior num teste paramétrico do que num teste não paramétrico. Assim, os testes não paramétricos devem apenas utilizar-se quando não existe alternativa, isto é, quando não é possível validar as condições de aplicação dos testes paramétricos, ou quando a escala de medida da variável dependente é qualitativa (situação na qual se podem utilizar os testes não paramétricos em detrimento dos testes paramétricos), [52].

Os testes paramétricos mais comummente usados em inferência estatística pressupõem a verificação simultânea das seguintes condições:

1. Que a variável dependente possua distribuição Normal;

2. As variâncias populacionais sejam homogéneas caso estejamos a comparar duas ou mais populações.

Seguidamente são apresentadas metodologias para testar a verificação das condições supracitadas.

3.2.1 Distribuição Normal

A distribuição Normal é o modelo probabilístico contínuo primordial, de importância indiscutível, visto que consiste numa suposição subjacente aos principais testes estatísticos, tais como teste t, análise de regressão linear, análise discriminante e analise da variância (ANOVA), [57]. Também inúmeros fenómenos aleatórios podem ser descritos de forma aproximada por este modelo. Assim, o desenvolvimento de análises estatísticas pressupõe a realização prévia de estudos relativos à normalidade dos dados. Quando a suposição da normalidade é violada, as interpretações e inferências de inúmeros testes estatísticos podem já não ser fidedignos ou válidos, [50].

Os três procedimentos comummente aplicados para avaliar se uma amostra aleatória de observações independentes de tamanho n provém de uma distribuição Normal são: métodos gráficos (histogramas, *boxplots*, *qqplots*), características sumárias (índices de *skewness* e *kurtosis*) e teste de normalidade formais (teste de Shapiro-Wilk, teste de Anderson-Darling, teste de Kolmogorov-Smirnov, entre outros), [75, 22].

Os *qqplots* e os histogramas constituem uma ferramenta primordial na medida em que permitem detetar visualmente a existência ou inexistência de normalidade dos dados, [52].

Diagramas quantil-quantil (diagramas *qqplot*) são representações gráficas dos quantis amostrais, calculados usando os dados da amostra original, em comparação com os quantis esperados. Assim, o conceito inerente ao *qqplot* é o de que, caso a variável em estudo siga determinada distribuição (como por exemplo a distribuição Normal), os quantis empíricos, isto é, os quantis calculados a partir de uma amostra, formam uma linha reta contra os quantis teóricos, os quais são calculados com base em estimativas dos parâmetros da distribuição (sendo que no caso da distribuição Normal é a média e o desvio padrão). Com efeito, o *qqplot* deve, idealmente, apresentar-se como a bissetriz dos quadrantes ímpares caso os dados sejam próximos da distribuição Normal, [46, 64].

A deteção de normalidade dos dados a partir da observação visual do histograma permite identificar grandes assimetrias, descontinuidades de dados e picos multimodais. Com efeito, recorre à existência de uma frequência mais alta no centro, seguida de um decrescimento gradual para as caudas de forma simétrica, assumindo a forma de sino, [37]. Verifica-se ainda a localização da média e da mediana próximas entre si (visto que devem ser aproximadamente iguais), localizando-se no centro do histograma (perto da moda).

Os testes formais de normalidade enunciados são testes que comparam a função de distribuição empírica que é estimada com base nos dados, com a função de distribuição cumulativa da distribuição Normal, e analisa se existe alguma relação entre elas. Esta classe de testes designa-se por testes EDF (*Empirical Distribution Function*). Dufour *et al.* (1998) descreveu os testes EDF como testes baseados numa medida de desfasamento entre as distribuições empíricas e hipotéticas, [22].

3.2.1.1 Teste de Shapiro-Wilk

O teste de Shapiro-Wilk foi proposto originalmente em 1965, tendo sido o primeiro teste capaz de detetar a normalidade de uma amostra. O teste permite obter resultados satisfatórios, tendo sido destacado pelas suas propriedades de referência, principalmente em amostras de tamanho inferior ou igual a 50 observações, [75, 44].

O teste de Shapiro-Wilk objetiva testar se os dados de uma amostra são provenientes de uma distribuição Normal. Dada uma amostra aleatória ordenada $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ a estatística de teste inerente é:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_{(i)} - \bar{X}\right)^2}, \tag{3.5}$$

onde:

- $X_{(i)}$ representa a i-ésima observação da variável X na amostra ordenada;
- $\bar{X}$ é a média da variável X na amostra;
- a_i são as constantes geradas a partir da média, variância e covariância de ordem n com a distribuição Normal. Os valores encontram-se tabelados. O cálculo de a_i =

$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$ e $m = (m_1, ..., m_n)^T$ são os valores esperados das estatísticas de ordem de uma amostra aleatória de variáveis independentes e identicamente distribuídas provenientes de uma distribuição Normal, e V a matriz de covariância das referidas estatísticas de ordem.

O valor de W encontra-se entre zero e um. Os valores críticos desta estatística de teste $(W_{(1-\alpha,\, n)})$ encontram-se tabelados, para cada nível de significância α e tamanho da amostra n, sendo que se rejeita H_0 ao nível de significância α quando $W_{obs} < W_{(1-\alpha,\, n)}$, [46, 55]. Desta forma, valores reduzidos de W induzem a rejeição da hipótese de Normalidade, enquanto que valores elevados desta estatística indicam a normalidade da amostra, [64].

O teste de Shapiro-Wilk foi inicialmente modificado por Royston em 1982, [55], de modo a alargar a sua utilização a amostras de dimensões superiores. Em 1999 Royston forneceu um novo algoritmo que tem vindo a ser utilizado, com bons resultados, para amostras com dimensão $3 \leq n \leq 5000$, [22, 64].

3.2.1.2 Teste de Kolmogorov-Smirnov

O presente teste é particularmente indicado para distribuições contínuas, uma vez que não se aplica a dados qualitativos nem a variáveis discretas. Pressupõe que os parâmetros da distribuição sejam pré-especificados, visto que não devem ser estimados a partir da amostra, [50, 75].

O teste de Kolmogorov-Smirnov (teste K-S) pertence à classe ''*supremum*' das estatísticas de EDF, sendo que esta classe de estatísticas se baseia na maior diferença vertical entre a distribuição hipotética e empírica, [15].

O teste K-S está relacionado com o conhecimento da função distribuição empírica $F_n(x)$. Assim, por forma a calcular a estatística do teste, é necessário começar por ordenar as observações da variável X por ordem crescente. Com efeito, dados n pontos ordenados $x_{(1)} < x_{(2)} < ... < x_{(n)}$, Conover (1999), [15], definiu a estatística de teste proposta por Kolmogorov (1933), como sendo

$$T = sup_x \left| F^*(x) - F_n(x) \right|, \tag{3.6}$$

onde *sup* designa o supremo, $F^*(x)$ é a função distribuição de probabilidades da distribuição hipotética e $F_n(x)$ é a função distribuição empírica da amostra a usar no teste. No teste de Kolmogorov-Smirnov, $F^*(x)$ representa a função distribuição da distribuição Normal, com média e desvio padrão conhecidos, μ e σ , respetivamente. A estatística de teste T apresentada representa a maior distancia (na vertical) entre a função de distribuição empírica e a função distribuição da população que se está a testar, [68].

O teste de Kolmogorov-Smirnov pretende testar se:

$$H_0 : \; F(x) = F^*(x) \; para \; todo \; x \in \mathbb{R}$$

(os dados/a amostra segue a distribuição especificada)

$$H_1 : \ \exists x \in \mathbb{R} : \ F(x) \neq F^*(x)$$

(os dados/a amostra não segue a distribuição especificada)

Rejeita-se H_0 quando a distribuição da amostra se afasta da distribuição hipotética, isto é, para valores elevados de $\left|F^*(x) - F_n(x)\right|$. Portanto, se o valor de T exceder o quantil $1-\alpha$ da distribuição da estatística, isto é, $T_{obs} \geq T_{(1-\alpha,\, n)}$, verifica-se que há motivos para rejeitar H_0, ao nível de significância α, [73].

Note-se que o teste de K-S apenas é considerado apropriado quando os parâmetros populacionais μ e σ da distribuição hipotética são totalmente conhecidos. Não obstante, tal situação é pouco comum. Assim, a estatística T não pode ser aplicada com rigor quando, em vez dos verdadeiros valores de μ e σ, se conhecem apenas estimativas amostrais. Para corrigir este problema, foi proposta por Lilliefors, em 1967, uma modificação deste teste, a qual se considera útil na comparação entre a distribuição de frequências acumuladas das observações da variável e a distribuição teórica cujos parâmetros foram estimados a partir da amostra. Neste caso, os parâmetros têm de ser estimados com base no próprio conjunto de dados com que é efetuado o teste, o que obrigou a recalcular a distribuição da estatística do teste. Ao teste KS assim adaptado designa-se por **Lilliefors K-S**,[52].

3.2.1.3 Teste de Anderson-Darling

De acordo com, Arshard *et al.*, [11], este teste é o mais poderoso dos testes EDF. A estatística do teste de Anderson-Darling (AD) pertence à classe quadrática da estatística EDF, na qual é baseada a diferença quadrática $\left(F_n(x) - F^*(x)\right)^2$. Anderson e Darling (1954) definiram a estatística de teste como:

$$W_n^2 = n \int_{-\infty}^{+\infty} \left[F_n(x) - F^*(x)\right]^2 \psi\left(F^*(x)\right) dF^*(x), \tag{3.7}$$

onde ψ é uma função de pesos não negativa que pode ser definida por

$$\psi = \left[F^*(x)(1 - F^*(x))\right]^{-1}$$

.

Por forma a facilitar o cálculo desta estatística, é possível escrever a equação (3.7) como sendo:

$$W_n^2 = -n - \frac{1}{n} \sum (2i-1) \left\{\log(F^*(X_i)) + \log(1 - F^*(X_{n+1-i}))\right\}, \tag{3.8}$$

onde $F^*(X_i)$ é a função distribuição cumulativa de uma distribuição especifica; X_i's são os dados ordenados $(i = 1, ..., n)$ e n designa o tamanho da amostra, [37, 22, 64].

3.2.2 Homogeneidade das variâncias

A suposição de homogeneidade das variâncias pressupõe que amostras independentes distintas possuem a mesma variabilidade, mesmo sendo provenientes de populações diferentes.

O referido pressuposto é exigido quando são aplicados inúmeros testes de hipóteses (como é o caso da ANOVA e o teste de *t-student*, por exemplo), [39].

A execução de um teste sem verificação prévia da homogeneidade das variâncias pode ter um impacto significativo nos resultados, podendo mesmo invalidá-los completamente. O enviesamento dos resultados depende do teste usado, e do quão sensível este é relativamente a variâncias desiguais. A assunção de variâncias iguais é ainda inerente à regressão linear, visto que a metodologia de regressão linear é tanto mais eficiente quanto menor a variância dos dados (visto que a dispersão dos dados é menor), [36, 52].

Existe uma panóplia de testes usados para avaliar a homogeneidade da variância numa amostra, entre os quais se destacam, o **teste de Levene**, o **teste de Fligner Killeen** e o **teste de Bartlett**. Os testes de Levene e Fligner Killeen são habitualmente usados quando a amostra não provém de uma distribuição Normal, enquanto que o teste de Bartlett é aplicado quando as amostras provém de uma população com distribuição Normal.

3.2.2.1 Teste de Levene

O **teste de Levene** consiste numa técnica robusta que, na ausência de normalidade, permite a obtenção de erros de rejeição da hipótese nula quando se verifica a sua veracidade, bastante reduzidos. É aplicável a diversas distribuições de probabilidade, [35].

Levene (1960) propôs uma estatística para testar a igualdade de variâncias para estudos balanceados, que, posteriormente, foi generalizada para estudos não balanceados. No presente teste, assume-se como hipótese nula a homogeneidade das variâncias, ou seja, pressupõe-se que a variabilidade das populações ou grupos em causa é a mesma, [35, 40].

Assim, suponha-se que são tomadas $k \geq 2$ amostras aleatórias independentes entre si, $X_{i1}, X_{i2}, ..., X_{in_i},\ i = 1, ..., k$. A amostra i representa uma coleção de n_i variáveis aleatórias independentes e identicamente distribuídas (iid), com distribuição G_i, com média μ_i e variância σ_i^2 desconhecidas. A hipótese nula de igualdade de variâncias, traduz-se por

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 , i = 1, ..., k,$$

a qual é testada contra a hipótese alternativa de que nem todas as variâncias são iguais, isto é,

$$H_1 : \sigma_i^2 \neq \sigma_j^2 , para\ algum\ i \neq j.$$

Denotem-se os desvios absolutos das variáveis X_{ij} em relação às médias ou às medianas amostrais dos grupos $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ por $Z_{ij} = \left| X_{ij} - \bar{X}_i \right| , j = 1, ..., n_i$ e $i = 1, ..., k$. Define-se a estatística de teste:

$$w = \left(\frac{n-k}{k-1} \right) \frac{\sum_{i=1}^{k} n_i (\bar{Z}_i - \bar{Z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2}, \tag{3.9}$$

onde:

- $\bar{Z}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$ é a média de Z_{ij} no grupo i;

- $\bar{Z}_{..} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Z}_i$ é a média de todos os Z_{ij};
- $n = \sum_{i=1}^{k} n_i$;
- k é o número de grupos diferentes aos quais as amostras pertencem;
- n_i é o número de casos no i-ésimo grupo;
- X_{ij} é o valor da variável de medida, do j-ésimo caso no i-ésimo grupo, que pode ser a média ou a mediana da amostra;
- $Z_{ij} = \left|X_{ij} - \bar{X}_i\right|,\ j = 1, ..., n_i \ e\ i = 1, ..., k.$

O teste de Levene consiste em rejeitar H_0 se $W_{obs} > F_{(1-\alpha),\,(k-1,n-k)}$.

Note-se que $F_{(1-\alpha),\,(k-1,n-k)}$ representa o quantil de ordem $1-\alpha$ da distribuição $F_{(k-1,n-k)}$ e α é o nível de significância do teste. Portanto, o teste é uma análise de variância com um fator na variável 'desvio absoluto', Z_{ij}. O uso de Z_{ij} em detrimento de Z_{ij}^2 faz com que o critério do teste se torne menos sensível à normalidade, [69, 45].

3.2.2.2 Teste de Fligner Killeen

O **teste de Fligner Killeen** objetiva, de forma análoga ao teste de Levene, avaliar a homogeneidade das variâncias de grupos com base em classificações. É comummente usado quando os dados não provêm de uma população com distribuição Normal, ou quando se verifica a existência de *outliers* na amostra, [41].

A estatística do teste de Fligner-Killeen é dada por:

$$F_0^2 = \frac{\sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2}{V^2}, \tag{3.10}$$

onde:

- $\bar{X}_j$ designa o valor da média da j-ésima amostra, donde se tem que

$$\bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{N,m_i},$$

  e x_{N,m_i} é o valor de pontuação crescente para a i-ésima observação na j-ésima amostra.

- $\bar{X}$ é o valor médio total de todos os $x_{N,i}$, ou seja,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_{N,i};$$

- V^2 é a variância da amostra, isto é,

$$V^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{N,i} - \bar{x})^2;$$

- $N = \sum_{j=1}^{k} n_j$.

Para amostras de tamanho elevado, o teste de Fligner-Killeen possui distribuição assintótica de Qui-Quadrado com $(k-1)$ graus de liberdade, ou seja, $F_0^2 \sim \chi^2_{(k-1)}$.

A hipótese nula do teste de Fligner-Killeen, à semelhança do teste de Levene, assume a homogeneidade das variâncias dos grupos, ou seja, H_0 é tal que:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \ , i = 1, ..., k.$$

A hipótese nula é testada contra a hipótese alternativa de que nem todas as variâncias são iguais, ou seja

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \, , para \ algum \ i \neq j,$$

onde σ_i^2 designa a variância de uma determinada variável no grupo ou amostra i.

A aplicação do presente teste recorrendo a *softwares* estatísticos, permite obter o *p-value* inerente ao teste. Assim, para um nível de significância α, se o *p-value* for inferior a α, a decisão consiste na rejeição da existência de homogeneidade das variâncias entre os grupos, [42, 45].

3.3 Testes não paramétricos

Os testes não paramétricos são usualmente aplicados quando pelo menos uma das condições de aplicação dos testes paramétricos não se verifica.

Este tipo de testes não exige o conhecimento da distribuição da variável em estudo, ou seja, as distribuições das populações envolvidas não necessitam de pertencer a uma família especifica de distribuições de probabilidade (Normal, Uniforme, Exponencial, entre outros). Desta forma, os testes não paramétricos podem ser também designados por "*distribution free tests*" (testes livres de distribuição). Isto é, os métodos não paramétricos usam procedimentos que são aplicáveis independentemente da distribuição da população. Por vezes, são apenas exigidas algumas hipóteses, como a simetria, a continuidade da distribuição, ou a ordenação dos dados, [73].

3.3.1 Teste de Kruskal-Wallis

O **teste de Kruskal-Wallis** é um teste não paramétrico comummente usado quando se pretende comparar mais do que duas amostras independentes, por forma a verificar se as amostras provêm da mesma população, [73]. Ou seja, permite detetar se, ao comparar mais do que duas amostras independentes, há diferença entre pelo menos duas, [68].

Contrariamente a outros testes estatísticos paramétricos (por exemplo, ANOVA *one way*), cuja análise de variância depende da hipótese de que todas as populações em confronto são independentes e normalmente distribuídas, o teste de Kruskal-Wallis pressupõe que as distribuições sejam contínuas e que difiram apenas na sua localização. Assim, o teste de Kruskal-Wallis pode surgir como alternativa não paramétrica à Anova *one way* paramétrica,[52, 68].

O teste de Kruskal-Wallis pode ser usado para testar se duas ou mais amostras provém de uma mesma população ou se, pelo menos, uma das populações é diferente ou se, de igual modo, as amostras provêm de populações com a mesma distribuição. Formalmente, as hipóteses inerentes ao teste podem escrever-se como:

$$H_0 : F(X_1) = F(X_2) = ... = F(X_k) \quad vs \quad H_1 : \exists i, j : F(X_i) \neq F(X_j),\ i \neq j,\ i, j = 1, ..., k.$$

As hipóteses indicadas matematicamente, traduzem-se pelo teste da hipótese nula de que as distribuições dos valores da variável dependente são idênticas nas k populações. Esta hipótese é testada contra a hipótese de que existe pelo menos uma população onde a distribuição da variável dependente é diferente de alguma das distribuições das outras populações sob estudo, [49].

Em particular, supondo-se que θ_i designa a mediana da população i $(i = 1, ..., k)$, onde k designa o número de grupos, as hipóteses do teste de Kruskal-Wallis são frequentemente escritas como:

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k \quad vs \quad H_1 : \exists i, j : \theta_i \neq \theta_j,\ i \neq j,\ i, j = 1, ..., k.$$

Assim, as hipóteses matemáticas apresentadas traduzem-se na suposição da igualdade das

medianas das k populações pela hipótese nula, contra a hipótese alternativa da assunção de que existe pelo menos um par de medianas significativamente diferentes, [49].

O cálculo da estatística de teste inicia-se com a ordenação crescente de todas as observações das diferentes amostras, sendo atribuída a cada uma a sua ordem na amostra global e mantendo a origem da observação. Assim, a estatística de teste é dada por:

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{N^3 - N}}, \tag{3.11}$$

onde $R_j = \sum_{i=1}^{n_j} r_{ij}$ representa a soma das ordens de cada uma das j ($j = 1, ..., k$) amostras e N ($N = n_1 + n_2 + ... + n_k$) é a dimensão da amostra global. O denominador designa a correção necessária caso existam mais de dois grupos de empates. O número de grupos é designado por g, enquanto que t designa o número de observações em cada grupo de empates. Sob H_0, $H \sim \chi^2_{1-\alpha,(k-1)}$ e, portanto, rejeita-se H_0 se $H_{obs} \geq \chi^2_{1-\alpha,(k-1)}$. O menor valor de α a partir do qual $H_{obs} \geq \chi^2_{1-\alpha,(k-1)}$, é a probabilidade de significância (*p-value)*, calculado a partir de programas estatísticos (rejeita-se H_0 se $p - value \leq \alpha$), [68, 73].

Salienta-se por fim que caso existam apenas duas amostras a comparar, o teste de Kruskal-Wallis é em tudo semelhante ao teste não paramétrico de Wilcoxon-Mann-Whitney.

3.3.2 Teste de Wilcoxon-Mann-Whitney

O **teste de Wilcoxon-Mann-Whitney**, ou simplesmente, Mann-Whitney, designa o teste não paramétrico adequado para comparar as funções de distribuição de uma variável medida entre duas amostras independentes. É comummente usado em situações onde se dispõe de amostras de tamanho reduzido ou muito diferente, cujas distribuições não são Normais ou são muito enviesadas e/ou não existe homogeneidade das variâncias, surgindo assim como alternativa ao teste-t, [49].

O teste de Wilcoxon-Mann-Whitney pressupõe a independência das amostras, que as distribuições sejam contínuas e que a variável de interesse seja passível de ser medida numa escala suscetível de ser ordenada, [52].

O presente teste pretende comparar as medidas de localização de duas amostras independentes, isto é, testar se as distribuições são iguais em localização, verificando se pertencem ou não à mesma população. Desta forma, o interesse reside em determinar se uma população tende a ter valores maiores do que a outra, ou se possuem a mesma mediana. As hipóteses estatísticas do teste de Mann-Whitney podem escrever-se como:

$$\begin{aligned}
&H_0 : F(X_1) = F(X_2) \quad vs \quad H_1 : F(X_1) \neq F(X_2);\\
&H_0 : F(X_1) = F(X_2) \quad vs \quad H_1 : F(X_1) > F(X_2);\\
&H_0 : F(X_1) = F(X_2) \quad vs \quad H_1 : F(X_1) < F(X_2).
\end{aligned}$$

As hipóteses do teste de Mann-Whitney são também, frequentemente, escritas em função das medianas populacionais. Suponha-se que θ_i designa a mediana da amostra i ($i = 1, ..., k$),

onde k designa o tamanho da população, [68, 73].

As hipóteses do teste, escritas em função das medianas populacionais, são:

$H_0 : \theta_1 = \theta_2 \quad vs \quad H_1 : \theta_1 \neq \theta_2;$

$H_0 : \theta_1 = \theta_2 \quad vs \quad H_1 : \theta_1 > \theta_2;$

$H_0 : \theta_1 = \theta_2 \quad vs \quad H_1 : \theta_1 < \theta_2.$

No entanto, neste caso, as conclusões são válidas apenas quando as funções de distribuição das duas populações sob estudo são idênticas. Note-se que se $F(X_1) = F(X_2)$ então tem-se que $\theta_1 = \theta_2$, mas que, se $\theta_1 = \theta_2$ isto não implica que $F(X_1) = F(X_2)$. Em termos práticos é mesmo possível ter duas amostras com a mesma mediana, mas ainda assim o teste de Mann-Whitney indicar diferenças estatisticamente significativas entre as duas populações. Não existe, neste tipo de conclusão, qualquer incoerência já que o teste é um teste às distribuições e apenas quando estas são idênticas, se pode extrapolar os resultados para as medianas populacionais, [52].

O teste baseia-se nas estatísticas de Wilcoxon (W) e de Mann-Whitney (U).

Assim, a estatística de Wilcoxon (W) é calculada da seguinte forma:

1. As $N = n_1 + n_2$ observações são ordenadas por ordem crescente. As ordens são representadas por r_{ij}, $i = 1, 2$ e $j = 1, ..., n_1$ ou n_2.

2. De seguida calcula-se a soma das ordens para cada uma das amostras:

$$R_1 = \sum_{j=1}^{n_1} r_{1j}, \tag{3.12}$$

$$R_2 = \sum_{j=1}^{n_2} r_{2j}. \tag{3.13}$$

   Note-se que a soma de todas as ordens é $N(N+1)$, pelo que R_2 também poderia ser escrito como sendo $R_2 = N(N+1) - R_1$.

3. A estatística de Wilcoxon é dada por:

$$W = \min(R_1, R_2). \tag{3.14}$$

A estatística de Mann-Whitney (U) é obtida de acordo com a seguinte regra:

É dada pelo número de vezes que uma observação da amostra 1 precede uma observação da amostra 2 (U_1) ou pelo número de vezes que uma observação da amostra 2 precede uma observação da amostra 1 (U_2), após a ordenação das N observações.

Assim, os valores de U_1 e U_2 são calculados da seguinte forma:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \tag{3.15}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = n_1 n_2 - U_1. \tag{3.16}$$

Desta forma, a estatística U de Mann-Whitney é dada por:

$$U = \min(U_1, U_2). \tag{3.17}$$

Os valores desta estatística encontram-se tabelados. Não obstante, pode recorrer-se à estatística de teste Z se N_1 e N_2 forem suficientemente elevados:

$$Z = \frac{U - E(U)}{S_W} \sim N(0,1), \tag{3.18}$$

onde $E(U) = \frac{N_1 N_2}{2}$ é o valor esperado da soma das ordens e $S_W = \sqrt{\frac{N_1 N_2 (N_1+N_2-1)}{12}}$ é o desvio padrão das ordens. Neste caso o *p-value* é calculado a partir da estatística de teste Z que possui distribuição assintótica $N(0,1)$.

3.3.3 ANOVA a dois fatores não paramétrico

Na Natureza, verifica-se que apenas um número reduzido de fenómenos podem ser explicados, ou são influenciados, apenas por um fator. A ANOVA consiste num método de destaque, pois permite estudar a influência de, pelo menos dois fatores, no comportamento da variável dependente (quantitativa). Assim, a presente metodologia permite estudar o efeito não só de cada um dos fatores, mas também estudar a possível influência que cada um dos fatores pode exercer sob a resposta da variável dependente e sob o outro fator. Este efeito designa-se por efeito de interação (ou moderação) entre os fatores. Assim, o desígnio da presente metodologia (ANOVA a dois fatores ou ANOVA *two-way*) decorre da possibilidade de estudar a influência de cada um dos fatores e o efeito da interação entre ambos na variável dependente, [31, 72].

O cariz da presente metodologia pode ser paramétrico ou não paramétrico, sendo que esta última é usada quando os pressupostos dos testes paramétricos não são satisfeitos, surgindo assim a ANOVA a dois fatores não paramétrico (também designada por ANOVA por *ranks*).

Suponha-se a existência de dois fatores: A e B. As observações da variável dependente (Y) podem organizar-se numa tabela de duas entradas (passível de ser consultada no Anexo 1). O fator A tem a níveis ($i = 1, ..., a$) e o fator B tem b níveis ($j = 1, ..., b$). Cada uma das combinações dos níveis do fator A e do fator B possui r repetições.

O modelo de ANOVA *two-way* pode escrever-se como:

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr}, \tag{3.19}$$

em que α_i representa o efeito do fator A, β_j o efeito do fator B, γ_{ij} a interação entre os fatores, e μ e ε_{ijr} são respetivamente a média global e os erros, $\varepsilon_{ijr} \sim N(0, \sigma)$.

Baseado no modelo teórico da ANOVA na população, é possível escrever o modelo da metodologia anterior, mas agora a partir das observações amostrais:

$$y_{ijr} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}) + (y_{ijr} - \bar{y}_{ij}), \tag{3.20}$$

onde:

- y_{ijr} representa a observação ijr;
- $\bar{y}$ é a média geral amostral (estimativa de μ);
- $(\bar{y}_i - \bar{y})$ é o efeito do fator A (estimativa de α_i);
- $(\bar{y}_j - \bar{y})$ é o efeito do fator B (estimativa de β_j);
- $(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})$ representa a interação entre fatores (estimativa de γ_{ij});
- $(y_{ijr} - \bar{y}_{ij})$ representa os resíduos (estimativa de ε_{ijr}).

Na ANOVA a dois fatores (não paramétrico), objetiva-se testar se as medianas para cada nível do fator A e do fator B são ou não iguais (se forem iguais então os fatores não têm um efeito significativo). Assim, as hipóteses para este teste traduzem-se por:

H_0^A: O fator A não tem um efeito significativo sobre a variável dependente (Y).

vs

H_1^A: O fator A tem um efeito significativo sobre a variável dependente (Y).

H_0^B: O fator B não tem um efeito significativo sobre a variável dependente (Y).

vs

H_1^B: O fator B tem um efeito significativo sobre a variável dependente (Y).

H_0^γ: Não existe interação entre os fatores A e B.

vs

H_1^γ: Existe interação entre os fatores A e B.

Para testar cada uma das hipóteses anteriores é necessário calcular uma estatística de teste H para o fator A, para o fator B, e para a interação $\gamma = A \times B$. O cálculo desta estatística começa por ordenar todas as observações Y_{ijk} por ordem crescente - ordens r_{ijk} - mantendo a identificação da origem da observação relativamente aos fatores em estudo (aos empates atribuem-se ordens médias), [52].

Em seguida calcula-se a soma dos quadrados das ordens para cada fator:

i) O valor de $SQOF_A$ é dado por:

$$SQOF_A = \frac{\sum_{i=1}^{a} \left(\sum_{j=1}^{b} \sum_{l=1}^{n} r_{ijl} \right)^2}{b \times n} - \frac{N(N+1)^2}{4}, \tag{3.21}$$

seguindo uma distribuição Qui-Quadrado, χ^2_{a-1}, $(g.l_A)$;

ii) O valor de $SQOF_B$ é dado por:

$$SQOF_B = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{l=1}^{n} r_{ijl}\right)^2}{a \times n} - \frac{N(N+1)^2}{4}, \tag{3.22}$$

que segue uma distribuição Qui-Quadrado, χ^2_{b-1} $(g.l_B)$;

iii) A soma dos quadrados das ordens das amostras (isto é, as combinações dos níveis dos fatores A e B), é dada por:

$$SQOA = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{l=1}^{n} r_{ijl}\right)^2}{n} - \frac{N(N+1)^2}{4}, \tag{3.23}$$

tendo uma distribuição Qui-Quadrado, χ^2_{ab-1};

iv) Finalmente, a soma das ordens das interações entre os fatores $(A \times B)$ é dada por:

$$SQO_{A\times B} = SQOA - SQOF_A - SQOF_B, \tag{3.24}$$

a qual segue uma distribuição Qui-Quadrado, χ^2, com $g.l_{A\times B} = g.l_A \times g.l_B = [(a-1)(b-1)]$;

v) A estatística de teste H define-se da seguinte forma:

- Para cada fator, $H_F = \frac{SQO_F}{QMOT}$.
- Para a interação entre os fatores, $H = \frac{SQO_{A\times B}}{QMOT}$.

vi) O $QMOT$ é o quadrado médio das ordens totais calculado pela fórmula:

$$QMOT = \frac{N(N+1)}{12}; \tag{3.25}$$

alternativamente, caso existam vários grupos de empates, o $QMOT$ deve ser calculado pela fórmula:

$$QMOT = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} r_{ijl}^2 - \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} r_{ijl}\right)^2}{N}}{N-1}. \tag{3.26}$$

Assim, os cálculos supracitados podem ser resumidos num quadro do tipo:

Tabela 3.1: Exemplo de organização de resultados da ANOVA a dois fatores.

Origem da variação	**Soma dos quadrados das ordens**	**Graus de liberdade**	**H**
Fator A	$SQOF_A$	$a-1$	$\frac{SQOF_A}{QMOT}$
Fator B	$SQOF_B$	$b-1$	$\frac{SQOF_B}{QMOT}$
Interação entre A e B	$SQOF_{A\times B}$	$(a-1)(b-1)$	$\frac{SQOF_{A\times B}}{QMOT}$
Total	SQOA	ab-1	

Sob H_0, H tem uma distribuição aproximada de Qui-Quadrado, χ^2, com os graus de liberdade da origem de variação a ser testada (isto é, com os graus de liberdade do fator A, fator B e/ou da interação entre A e B). Finalmente, rejeita-se H_0 com uma probabilidade de erro do tipo I não superior a α se $H_{obs} \geq \chi^2_{1-\alpha,(g.l)}$, [52].

Salienta-se, por fim, que segundo, [65], os resultados obtidos por aplicação de ANOVA a dois fatores não paramétrico podem não ser fidedignos, em situações em que os *p-values* obtidos relativamente aos dois fatores, implicam a rejeição das hipóteses nulas de que cada um dos fatores, A e B, não influenciam a variável dependente. Nestas situações o *p-value* do teste de interação dos dois fatores deflaciona artificialmente. Isto é, caso os resultados da ANOVA a dois fatores obtidos permitam evidenciar a existência de influencia de cada um dos fatores, isoladamente, na variável dependente, e caso o *p-value* relativo à interação implique também a rejeição da hipótese nula, que assume a influencia da interação nula dos fatores, tal pode não traduzir efetivamente a realidade. Assim, caso os resultados de ANOVA a dois fatores não paramétrico não levem à tripla rejeição, é plausível intuir que a metodologia aplicada aos dados é fidedigna. Caso contrário, deve-se suspeitar dos resultados obtidos.

3.4 Regressão

A análise de regressão consiste numa metodologia estatística amplamente usada em análise de dados, [39]. Os modelos de regressão são usados com diferentes objetivos: estudo do efeito causal de uma variável (ou um conjunto de variáveis) numa variável dependente, e, consequentemente, tornam possível a construção de uma função matemática que descreve essa relação; ou, podem ainda ser usados para prever o valor da variável de resposta, [52, 62, 31].

3.4.1 Regressão linear múltipla

A regressão linear múltipla designa uma técnica de análise multivariada, cujo propósito é estabelecer uma relação entre uma (ou mais) variáveis dependente(s) (ou resposta) e duas (ou mais) variáveis independentes (ou explicativas), permitindo explicar essa relação e, assim, analisar de que forma as variáveis explicativas são significativas para justificar a variabilidade da variável resposta.

Em regressão linear simples, é construído um modelo de regressão entre uma variável de resposta Y, e uma única variável explicativa x, cuja representação gráfica consiste numa função linear, [9]. O modelo de regressão linear múltipla representa uma generalização dessa metodologia, permitindo a inclusão de múltiplas variáveis explicativas ou preditoras, sendo que graficamente representa um plano hipergeométrico. Num modelo de regressão linear múltipla podem ser consideradas uma ou mais variáveis dependentes, em função do cariz inerente ao problema em estudo, [39, 45]. Assim, a um modelo de regressão onde seja considerada uma variável dependente e múltiplas variáveis independentes, designa-se por modelo de regressão linear múltiplo univariado. Por outro lado, a um modelo de regressão onde sejam consideradas múltiplas variáveis dependentes e múltiplas variáveis independentes, designa-se por modelo de regressão linear múltiplo multivariado, [36].

O modelo de regressão linear múltipla que relaciona n observações independentes em Y, é descrito pela i-ésima variável dependente, Y_i, modelada em função de um conjunto de variáveis independentes, $X_1, X_2, ..., X_K$. Tendo em conta que um modelo não é capaz de prever exatamente os valores observados, Y, é necessário introduzir uma variável aleatória ε associado à aplicação do modelo aos dados, que designa o erro ou distúrbio, [72, 62]. Assim, o modelo pode ser escrito por:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_K X_{iK} + \varepsilon_i \\ &= \sum_{j=0}^{K} \beta_j X_{ij} + \varepsilon_i \\ &= \mathbf{X_i^T}\beta + \varepsilon_i, \quad X_{i0} = 1. \end{aligned} \tag{3.27}$$

A expressão apresentada descreve a forma como a variável dependente Y_i é decomposta numa função linear de variáveis independentes X_{ij}, com $i = 1, ..., n$ e $j = 1, ..., K$. As variáveis independentes encontram-se relacionadas com os coeficientes de regressão, β_j, os quais simbolizam os declives parciais, e β_0 designa a ordenada na origem, [62]. Com efeito, os coeficientes de regressão $\beta_1, \beta_2, ..., \beta_K$ traduzem a quantidade de variação da variável resposta

associada à alteração de uma unidade da correspondente variável explicativa, condicionalmente à constância das restantes variáveis explicativas do modelo de regressão, [39, 5].

O modelo (3.27) assume ainda a existência de um termo aleatório, ε_i que designa o termo dos erros. Tendo em conta que ε_i é intrinsecamente não observável, é necessário simplificar os pressupostos inerentes aos erros do modelo. Assim, uma suposição chave que produz a identificação dos parâmetros desconhecidos da equação (3.27) é a independência entre os erros e as variáveis explicativas. São ainda assumidos outros pressupostos por forma a incrementar a eficiência do modelo. Com efeito, assume-se que os erros ε_i são independentes entre si, e identicamente distribuídos (i.i.d.), [20, 47]. O pressuposto de independência implica que a correlação dos erros ε entre um par de observações é zero, enquanto que o pressuposto da distribuição idêntica assume variância σ_ε^2 comum (isto é, homeoscedasticidade das variâncias dos erros). Assim, supõem-se que os termos dos erros do modelo são normalmente distribuídos, com média zero e variância σ_ε^2. Tal implica que, para os valores das variáveis explicativas, a variável de resposta deve ser normalmente distribuída, com média correspondente a uma função linear das variáveis explicativas e variância que não depende das referidas variáveis, [31]. Consequentemente, uma forma equivalente de escrever o modelo de regressão linear múltipla é $Y|X_1, ..., X_K \sim N(\mu, \sigma_\varepsilon^2)$, onde $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K$.

Pressupõe-se ainda a ortogonalidade das variáveis explicativas, isto é, as referidas variáveis devem ter valores de correlação nula ou próxima de zero, [47]. As conjeturas do modelo de regressão, e respetiva validação, serão exploradas com maior detalhe na secção 3.4.5.

Tendo por base o pressuposto de que os erros são i.i.d., é possível estimar a equação (3.27) fazendo uso do método dos mínimos quadrados, o qual é descrito na secção 3.4.2.1. Contudo, é possível usar o método dos mínimos quadrados na estimação dos coeficientes de regressão β_i caso não seja verificado o pressuposto dos erros serem i.i.d., pois o estimador é ainda consistente caso os ε sejam não correlacionados com as variáveis explicativas x, o que permite a sua convergência para o vetor do parâmetro, quando o tamanho do conjunto de dados é elevado, [39, 45].

Assim, a construção do modelo (3.27) inicia-se com a estimação dos coeficientes de regressão do modelo.

3.4.2 Estimação dos coeficientes de regressão

O principal objetivo inerente ao modelo de regressão linear é obter um conjunto de coeficientes de regressão. Desta forma, a construção do modelo de regressão descrito inicia-se com a estimação dos coeficientes de regressão fazendo uso de um método estatístico apropriado, tendo em conta, em particular, a presença no modelo do termo de perturbação aleatória. Com efeito, o principal objetivo inerente à estimação dos coeficientes do modelo de regressão consiste na obtenção de um modelo capaz de prever a variável dependente de forma tão próxima quanto possível dos valores reais da referida variável, [52].

Existe uma panóplia de metodologias estatísticas capazes de estimar os coeficientes do modelo de regressão, mormente, o método do Mínimos Quadrados e estimação pelo método da Máxima Verosimilhança. Na presente investigação ir-se-á implementar regressão linear múltipla recorrendo à estimação dos coeficientes de regressão pelo método do mínimos quadrados, pelo que seguidamente são explorados os conceitos teóricos inerentes ao mesmo.

3.4.2.1 O método dos mínimos quadrados

Considerando uma variável aleatória Y medida continuamente, um objetivo inerente ao modelo de regressão consiste na estimação da média populacional de Y condicional a um conjunto de K variáveis independentes a partir de uma amostra formada por n observações e averiguar de que forma o valor de Y varia com as variáveis explicativas $X_1, X_2, ..., X_K$, ou, de forma equivalente, determinar o efeito de X_j $(j = 1, ..., K)$ na variável independente Y, [72, 39]. Assim, a relação existente entre Y e as variáveis explicativas é modelada, de forma linear, com base em (3.27). O método dos mínimos quadrados consiste numa metodologia pautada de simplicidade capaz de estimar de forma ótima os coeficientes de regressão, β_j, com $j = 0, 1, ..., K$. Os valores ótimos são os que minimizam a soma do quadrado dos desvios em torno da média condicional, [52, 45]. Com efeito, tendo por base o modelo de regressão descrito na equação (3.27), o objetivo do método dos mínimos quadrados consiste em encontrar as estimativas de β que minimizem a soma dos desvios dos quadrados dos resíduos (ou seja, estimativas dos erros), isto é,

$$S(\beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - ... - \hat{\beta}_K X_{iK})^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{K} \hat{\beta}_j X_{ij} \right)^2 . \quad (3.28)$$

A minimização de $S(\beta)$ é calculada recorrendo à derivada parcial de $S(\beta)$, relativamente a β_j, ou seja,

$$\frac{\partial S(\beta)}{\partial S(\beta_j)} = -2 \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{K} \hat{\beta}_j X_{ij} \right) = 0, \quad j = 0, ..., K. \quad (3.29)$$

A derivada parcial descreve a taxa de variação da função da soma dos quadrados dos resíduos para quaisquer valores de β estimados, [9]. Assim, a soma dos quadrados dos resíduos é mínima quando a sua taxa de variação é nula. O cálculo do valor da equação (3.29) permite obter a expressão do cálculo das estimativas dos coeficientes do modelo de regressão, dada, em termos matriciais, por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad (3.30)$$

onde $\hat{\beta}$ designa o vetor dos coeficientes de regressão estimados, X representa a matriz $(n \times (K+1))$ de variáveis independentes, sendo X^T a matriz transposta de X, e y é o vetor $(n \times 1)$ dos valores da variável dependente do modelo.

O teorema de Gauss-Markov estabelece que $\hat{\beta}$ da equação (3.30) é o estimador mais eficiente entre os estimadores lineares não enviesados (estimadores BLUE), [59], com variância e covariância dadas por:

$$var(\hat{\beta}) = \sigma_{\varepsilon}^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (3.31)$$

Note-se que $var(\hat{\beta})$ é uma matriz $(K+1) \times (K+1)$, cujos elementos diagonais correspondem à variância dos estimadores, e os elementos localizados nas restantes entradas da matriz correspondem às covariâncias entre os estimadores, [31].

3.4.3 Métodos de seleção de preditores

Num problema de regressão linear múltipla, é possível conhecer à partida quais as variáveis independentes a incluir no modelo. Não obstante, principalmente em fases exploratórias da análise em regressão linear, as variáveis que influenciam a variável dependente podem ser desconhecidas, isto é, as variáveis que conduzem ao "melhor modelo" podem ser desconhecidas, e, consequentemente, a decisão da escolha pode ser complicada pela presença de colinearidade e dos seus efeitos sobre a magnitude e o sinal dos coeficientes de regressão. Assim, existem inúmeros métodos que auxiliam na seleção de variáveis exploratórias a considerar num modelo de regressão, e consequentemente, auxiliam na busca do "melhor modelo" de regressão. Note-se, porém, que nenhum deles garantidamente, conduz ao modelo "ótimo", tendo em conta as características inerentes aos conjuntos de dados, [52, 79].

a) **Seleção *Forward***:

Neste método de seleção de variáveis, o modelo inicial é formado apenas pela constante β_0. As variáveis devem encontrar-se ordenadas de forma decrescente dos seus valores de correlação (em valor absoluto) com a variável dependente Y. Assim, no primeiro passo introduz-se a variável que se encontra posicionada no topo da lista, e são calculados os valores do teste F de significância global do modelo e o valor do coeficiente de determinação múltiplo, R^2. Esta variável é adicionada ao modelo se a estatística F e o valor de R^2 forem incrementados. Seguidamente ir-se-á adicionar ao modelo construído a variável seguinte da lista ordenada, e assim sucessivamente, até que a adição de uma nova variável não permita incrementar os valores de F e de R^2, [52].

b) **Seleção *Backward***:

Neste método, o modelo é iniciado com a totalidade das K variáveis independentes, e no passo seguinte é calculada uma estatística F parcial e o coeficiente de determinação múltiplo, R^2, para cada variável, como se esta fosse a última a entrar no modelo. Seleciona-se a variável cuja eliminação permitir incrementar os valores de F e de R^2. Assim, no próximo passo o modelo de regressão é formado com menos uma variável. O procedimento repete-se agora para este último modelo construído com menos uma variável que o primeiro, e assim sucessivamente. O procedimento prossegue até que não existam variáveis no modelo, ou até ser construído o modelo que maximiza os valores da estatística F e de R^2, [23].

c) **Seleção *Stepwise***:

Este método conjuga as características inerentes aos dois métodos de seleção anteriores. Com efeito, no primeiro passo de seleção *Stepwise*, o modelo inicia-se com uma só variável independente (à semelhança do modelo de seleção *Forward*), mas a significância de cada adição é testada como no método *Backward*. Ou seja, em cada passo adiciona-se a variável cujo valor de correlação com a variável dependente Y é superior (em termos absolutos) e verifica-se se a eliminação de alguma das variáveis já incluídas no modelo permite incrementar os valores da estatística F e de R^2. São construídos inúmeros modelos de regressão, sendo que é selecionado o modelo que maximiza os valores da estatística F e de R^2. A vantagem inerente a este método é que permite a eliminação de variáveis previamente incluídas caso a adição de uma nova variável induza a redução da sua significância no modelo de regressão. Este procedimento termina quando nenhuma das variáveis independentes não incluídas

ainda no modelo, consegue ser inserida no modelo, com base na análise dos respetivos valores da estatística F e de R^2. Este método é substancialmente adequado quando existem correlações significativas entre as variáveis independentes, [79].

3.4.4 Inferência sobre o modelo de regressão linear múltipla

Após a determinação das estimativas $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$, é possível proceder à avaliação da influência quantitativa das variáveis independentes sobre a variável dependente na amostra. Assim, os critérios usados na seleção de variáveis e na avaliação do modelo de regressão construído incluem o teste de significância global de regressão, o teste de significância dos regressores e o cálculo do coeficiente de determinação R^2.

3.4.4.1 Teste de significância global de regressão

Considerando o modelo de regressão linear da equação (3.27), é plausível intuir que, se todos os coeficientes associados às variáveis explicativas do modelo forem simultaneamente iguais a zero, as variáveis explicativas seriam, no seu conjunto, consideradas estatisticamente não significativas para explicar a variável dependente, Y, [45].

Assim, para testar a significância global de regressão, ou seja, para testar as hipóteses

$$H_0: \ \beta_1 = \beta_2 = ... = \beta_K = 0 \quad vs \quad H_1: \ \exists\, \beta_j \neq 0, \quad j = 1, 2, ..., K$$

faz-se uso da seguinte estatística:

$$F = \frac{SQE/(K-1)}{SQR/(n-K)} = \frac{R^2/(K-1)}{R^2/(n-K)} \sim F_{(K-1,n-K)}, \tag{3.32}$$

onde $SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ é a soma de quadrados explicada e $SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ é a soma de quadrados dos resíduos de regressão. A estatística apresentada segue a distribuição F de Fisher-Snedecor, com $K-1$ graus de liberdade no numerador, e $n-k$ graus de liberdade no denominador, onde n e K representam, o número de observações do conjunto de dados e a quantidade de variáveis explicativas usadas no modelo de regressão, respetivamente.

Para um dado nível de significância, rejeita-se H_0 se $F_{obs} > F_{1-\alpha,(K-1,n-K)}$, e conclui-se pela significância global de regressão, ou seja, que as variáveis explicativas conjuntamente se relacionam de forma estatisticamente significativa com a variável dependente. Caso contrário, se H_0 não é rejeitada, conclui-se, para um determinado nível de significância, que a regressão não é globalmente significativa para explicar a variável dependente, [27, 80].

Salienta-se, por fim, que o teste é valido apenas para modelos com termo independente.

3.4.4.2 Teste de significância dos regressores

É ainda possível averiguar a significância dos regressores para o modelo. Para cada variável de resposta Y_i, o coeficiente β_j traduz a contribuição do regressor X_{ij} para explicar a variação em y_i, com $i = 1, ..., n$ e $j = 1, .., K$. Assim, caso β_j assuma valores aproximadamente nulos,

conclui-se que a contribuição não é preponderante. Consequentemente, é premente testar a significância de cada regressor na variável de resposta Y_i, o que se traduz nas seguintes hipóteses:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0, \quad j = 1, ..., K.$$

A estatística de teste associada ao presente teste é:

$$T_j = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}, \tag{3.33}$$

onde $s(\hat{\beta}_j)$ designa o erro padrão de cada estimador individual $\hat{\beta}_j$, para $j = 1, ..., K$.

Assim, considerando um nível de significância α, rejeita-se a hipótese nula quando o valor observado da estatística de teste T_j for significativamente diferente de zero, isto é, encontra-se fora do intervalo $\left(-t_{1-\frac{\alpha}{2},n-K-1}, t_{1-\frac{\alpha}{2},n-K-1}\right)$. A rejeição da hipótese nula, permite concluir que o regressor X_{ij} contribui significativamente para explicar a variação em y_i, na presença das restantes variáveis explicativas do modelo. Assim, salienta-se que o teste de significância dos regressores testa a significância de uma variável independente na presença das restantes variáveis do modelo, sendo que a remoção ou adição de variáveis pode implicar a alteração do resultado do teste.

3.4.4.3 Coeficiente de determinação múltiplo

O coeficiente de determinação múltiplo, R^2, consiste numa medida referente ao poder explicativo do modelo aplicado, isto é, faz referência à proporção de variação da variável dependente que é explicada pela(s) variável(eis) independente(s). Com efeito, quanto mais próximo de um for o valor de R^2, maior a quantidade de variação da variável independente que é explicada pela(s) variável(eis) independente(s). Por outro lado, um valor de R^2 próximo de zero induz a inadequação do modelo aplicado, [26, 45].

Note-se que este coeficiente não deve ser usado para comparar modelos que diferem relativamente ao número de variáveis independentes, uma vez que, de uma forma geral, a incorporação de mais uma variável independente tende a aumentar o valor de R^2, mesmo que a sua influencia na variável dependente seja reduzida. De forma alternativa, o coeficiente de R^2 ajustado é mais adequado, visto que geralmente aumenta se a adição da nova variável produzir um melhor ajustamento do modelo aos dados. Este coeficiente é dado por

$$R^2_{ajustado} = 1 - \frac{QME}{QMT},$$

onde QME designa o quadrado médio dos resíduos e $QMT = \frac{SQT}{n-1}$, com SQT a designar a variabilidade total de Y e n o número de variáveis independentes do modelo. Assim, é plausível intuir que o coeficiente ajustado apenas aumenta quando a variabilidade dos erros diminui relativamente à variabilidade total, [59].

3.4.5 Validação dos pressupostos de regressão linear múltipla

O modelo de regressão linear múltipla só é considerado adequado para fins de estimação e inferência de relações funcionais entre a variável dependente e as variáveis independentes, se um conjunto de pressupostos inerentes ao modelo forem válidos, [45].

Desta forma, após a estimação dos coeficientes, a regressão linear múltipla prossegue com com a validação dos pressupostos associados aos resíduos e à ortogonalidade entre as variáveis independentes.

3.4.5.1 A análise dos resíduos

A definição do modelo de regressão linear múltipla, através da equação (3.27), assume a existência de uma componente aleatória, designada por ε_i, que representa os erros aos quais foram impostas determinadas suposições,tal como explanado anteriormente. Tendo em conta que os valores dos erros são desconhecidos, essas hipóteses serão analisadas através da análise dos resíduos do modelo de regressão $\hat{\varepsilon}_i$, definidos pela diferença entre os valores observados ou reais e os valores estimados pelo modelo de regressão. Com efeito, a validação do modelo de regressão linear requer que se verifiquem, concomitantemente, as condições exploradas subsequentemente, [20, 39].

i) **Normalidade dos resíduos**

Os resíduos do modelo de regressão devem seguir uma distribuição Normal, de média nula e variância constante, isto é, $\hat{\varepsilon}_i \sim N(0, \sigma_\varepsilon^2)$. A normalidade dos resíduos pode ser verificada através de métodos gráficos, através de um *QQ-plot* dos resíduos, em que as observações se devem aproximar da bissetriz dos quadrantes ímpares (se os dados não estiverem estandardizados), ou usando testes de normalidade (abordados na secção 3.2.1), [50].

A violação do pressuposto de normalidade dos resíduos impede o cálculo dos intervalos de confiança das previsões obtidas com base no modelo de regressão linear, [72].

ii) **Autocorrelação dos resíduos**

No modelo de regressão linear múltiplo é postulada a ausência de autocorrelação entre as perturbações aleatórias, estabelecendo-se assim que é nula a covariância existente entre duas quaisquer perturbações aleatórias, pelo que a hipótese nula de autocorrelação se traduz por:

$$H_0 : \ Cov(\varepsilon_i, \varepsilon_j) = 0, \ \ \forall i, j = 1, 2, ..., n, \ \ i \neq j.$$

Estabelece-se a violação desta hipótese, ou seja, existirá correlação das perturbações aleatórias, desde que haja pelo menos dois erros cuja covariância seja diferente de zero, isto é, a hipótese alternativa do pressuposto de autocorrelação traduz-se por:

$$H_1 : \ \exists i, j, \ i \neq j : \ Cov(\varepsilon_i, \varepsilon_j) \neq 0.$$

Com efeito, o pressuposto da autocorrelação dos resíduos estabelece que a magnitude de um erro não deve influenciar a magnitude do erro seguinte, [45].

Assim, a validação deste pressuposto pode ser desenvolvido recorrendo a um procedimento gráfico, no qual são relacionados os valores preditos, $\hat{Y}_i$, com os respetivos resíduos, $\hat{\varepsilon}_i$. Idealmente, a inexistência de autocorrelação nos resíduos, deteta-se pela distribuição aleatória dos pontos entre os resíduos e os valores preditos, em torno da reta horizontal $\hat{\varepsilon}_i = 0$, situação descrita na figura 3.1, [52].

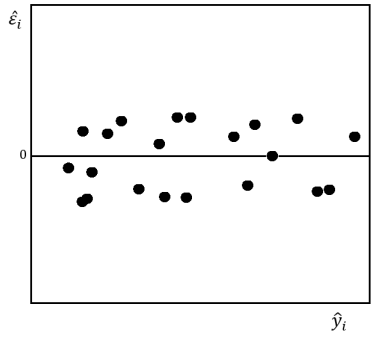


Figura 3.1: Gráfico entre os valores preditos e os resíduos do modelo de regressão linear, que permite detetar a inexistência de autocorrelação nos resíduos.

A análise da autocorrelação (dependência) dos resíduos de um modelo de regressão linear pode ainda recorrer a testes de autocorrelação dos resíduos, como são exemplo os testes de **Durbin-Watson**, **Box-Pierce** e **Ljung-Box**.

O trabalho conjunto de dois autores, Durbin e Watson (1971), permitiu a construção do **teste de Durbin-Watson**, o qual se baseia numa estatística, designada pela letra d ou pelas iniciais DW, definida da seguinte forma:

$$d = DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \approx 2(1 - \hat{\rho}_{\hat{\varepsilon}_t,\hat{\varepsilon}_{t-1}}) = 2 - 2\hat{\rho}_{\hat{\varepsilon}_t,\hat{\varepsilon}_{t-1}}, \tag{3.34}$$

em que $\hat{\varepsilon}_t$ se refere ao resíduo inerente à aplicação do método de estimação dos mínimos quadrados e pressupondo ainda que a autocorrelação seja gerada por um processo AR(1). O numerador traduz a soma de quadrados das diferenças entre resíduos de observações consecutivas, enquanto o denominador representa a soma de quadrados dos resíduos. Assim, tendo por base as características explanadas, a estatística de teste assume apenas valores positivos.

Com base na equação (3.34), é plausível afirmar que a estatística de teste DW assume valores entre 0 e 4. Assim, por um lado, caso $DW \approx 2$, conclui-se a rejeição da hipótese nula, e consequentemente da existência de autocorrelação entre os resíduos, na medida em que $\hat{\rho}_{\varepsilon_t,\varepsilon_{t-1}} \approx 0$. A hipótese alternativa do teste é escolhida com base no valor da estatística de teste DW. Desta forma, se $0 < DW < 2$, visto que, $0 < \hat{\rho}_{\varepsilon_t,\varepsilon_{t-1}} < 1$, a hipótese alternativa testa a exitência de autocorrelação positiva entre os resíduos. Por outro lado, se $2 < DW < 4$, tendo em conta que $-1 < \hat{\rho}_{\varepsilon_t,\varepsilon_{t-1}} < 0$ a hipótese alternativa testa a existência de autocorrelação negativa entre os resíduos, [45]. O teste de Durbin-Watson é bastante limitado visto que apenas permite que a autocorrelação seja gerada por processos AR(1).

Desta forma, o **teste de Breusch-Godfrey** surge como alternativa mais geral ao primeiro, permitindo colmatar a referida falha. A estatística de teste inerente ao mesmo é dada por:

$$BG = N \times R^2,$$

onde N designa o número de observações do conjunto de dados, e R^2 é o coeficiente de determinação múltiplo obtido por aplicação de regressão linear. A estatística de teste BG segue uma distribuição Qui-quadrado com K graus de liberdade, $\chi^2_{(K)}$, pelo que se rejeita a inexistência de autocorrelação nos resíduos caso $BG_{obs} > \chi^2_{(1-\alpha,K)}$, considerando um nível de significância α, [20, 62].

Alternativamente, destacam-se os **testes de Box-Pierce**, e de **Ljung-Box**, sendo que o segundo resultou de um melhoramento do primeiro, tornando-o mais generalizado. Ambos os testes assumem a hipótese nula de independência dos resíduos. [9]

As estatísticas de teste para o teste de Box-Pierce e para o teste de Ljung-Box são, respetivamente:

$$Q(k)_{Box-Pierce} = n \sum_{j=1}^{K} \hat{\rho}_j^2, \tag{3.35}$$

$$Q(k)_{Ljung-Box} = n(n-2) \sum_{j=1}^{K} \frac{\hat{\rho}_j^2}{n-j}, \tag{3.36}$$

onde k representa o número de desfasamentos na função de autocorrelação estimada, $\hat{p}$, e n é o tamanho da amostra em estudo.

Assim, no caso da estatística de **teste de Box-Pierce**, tem-se que $Q(k)_{Box-Pierce}$ terá aproximadamente uma distribuição Qui-quadrado com K graus de liberdade, $\chi^2_{(K)}$, enquanto que a estatística de **teste de Ljung-Box**, $Q(k)_{Ljung-Box}$, terá aproximadamente uma distribuição Qui-Quadrado com $(K-p-q)$ graus de liberdade, $\chi^2_{(K-p-q)}$, onde p e q representam as ordens do modelo ajustado. De realçar que os testes supracitados são aplicados aos resíduos de um modelo ARMA(p,q) ajustado, e não à série original, sendo que em tais aplicações a hipótese sob teste é que os resíduos do modelo ARMA(p,q) não são autocorrelacionados. Assim, os graus de liberdade p e q designam as ordem do modelo ARMA(p,q).

Portanto, no caso do **teste de Box-Pierce** rejeita-se a hipótese nula se $Q(k)_{Box-Pierce} > \chi^2_{(1-\alpha,K)}$, e no caso do **teste de Ljung-Box** rejeita-se a hipótese nula se $Q(k)_{Ljung-Box} > \chi^2_{(1-\alpha,K-p-q)}$, a um nível de significância α, [39].

Quando existe autocorrelação dos resíduos os estimadores do método dos mínimos quadrados serão não enviesados; contudo, os testes de significância e o cálculo de intervalos de confiança inerentes ao modelo de regressão construído são inválidos.

iii) **Homogeneidade das variâncias dos resíduos**

No modelo de regressão linear múltipla, assume-se ainda que a variável dependente possui média diferente em função das variáveis explicativas do modelo de regressão, porém, deve

possuir variância constante, independentemente dos valores das referidas variáveis. Tal consiste numa consequência da hipótese de homogeneidade (ou homoscedasticidade), que postula que os erros do modelo de regressão linear possuem igual variância, [45, 47].

A existência de heterocedasticidade nos erros do modelo afeta negativamente a estimação e inferência estatística inerente ao modelo de regressão linear múltipla construído, [39].

Neste contexto a hipótese de homoscedasticidade (igual variância) é dada por:

$$Var(\varepsilon|X_{2i}, X_{3i}, ..., X_{Ki}) = \sigma_\varepsilon^2, \ \forall i, \ 0 < \sigma_\varepsilon^2 < \infty,$$

a qual mais sinteticamente pode ser escrita por:

$$Var(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall i, \quad 0 < \sigma_\varepsilon^2 < \infty$$

A hipótese supracitada postula que a variância dos erros ou resíduos assume um valor constante, idêntica qualquer que seja a sequência ordenada $(X_{2i}, X_{3i}, ..., X_{Ki})$. Com efeito, para uma mesma sequência de variáveis explicativas, ter-se-ão diversos valores de Y, e, por conseguinte, inúmeros valores de ε; não obstante, a dispersão desses valores deve manter-se constante, [31].

Assim, a violação da hipótese de homoscedasticidade dos erros, [45], traduz-se por:

$$\exists^1(i, j) : \ Var(\varepsilon|X_{2i}, X_{3i}, ..., X_{ki}) \neq Var(\varepsilon|X_{2j}, X_{3j}, ..., X_{kj}),$$

a qual mais sinteticamente pode ser escrita como:

$$\exists^1(i, j) : \ Var(\varepsilon_i) \neq Var(\varepsilon_j)$$

A deteção de heterocedasticidade nos erros é desenvolvida com recurso a representações gráficas entre os resíduos e os valores ajustados, passíveis de serem consultados nas figuras 1a e 2 do Anexo B.2. As figuras apresentadas permitem evidenciar que a presença de heterocedasticidade nos resíduos é detetada pela existência de um 'V' horizontal dos pontos de dispersão dos resíduos pelos valores ajustados. Uma alternativa que permite consolidar as conclusões indagadas graficamente, recorre a testes estatísticos. Os testes estatísticos comummente usados para esse efeito são: o **teste de Breuch-Pagan** e o **teste de Koenker**.

O teste de Breusch-Pagan é comummente usado para testar a hipótese nula de que as variâncias dos erros são iguais (homocedasticidade), sendo indicado para amostras de elevada dimensão.

Assim, o pressuposto inerente ao **teste de Breusch-Pagan** é o de que os erros ε_i são independentes, com média nula e variância dada por:

$$Var(\varepsilon_i) = \sigma_i^2 = h(\alpha_1 + \alpha_2 Z_{i2} + \alpha_3 Z_{i3} + ... + \alpha_p Z_{ip}), \quad i = 1, ..., n,$$

ou recorrendo à forma matricial, é possível reescrever a expressão anterior como sendo:

$$Var(\varepsilon_i) = \sigma_\varepsilon^2 = h(\mathbf{Z_i}\alpha),$$

onde $h(.)$ é uma função que não é necessário especificar, de uma combinação linear de variáveis observáveis $Z_2, Z_3, ..., Z_p$, [52, 45]. Note-se que as variáveis Z_{ip} são as variáveis que se supõe estar na origem da heterocedasticidade.

Na existência de homoscedasticidadade, tem-se que $\alpha_2 = \alpha_3 = ... = \alpha_p = 0$, de forma que $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = h(\alpha_1)$, ou seja, uma constante. Se, pelo contrário, algum dos coeficientes $\alpha_2, \alpha_3, ..., \alpha_p$ for diferente de 0, então σ_i^2 dependerá da i-ésima observação, do que constitui uma violação da homoscedasticidadade dos resíduos.

Assim, a hipótese nula do teste de Breusch-Pagan é definida como sendo:

$$H_0 : \alpha_2 = \alpha_3 = ... = \alpha_p = 0.$$

A estatística de teste inerente ao teste de Breuch-Pagan é dada por:

$$BP = \frac{SQE}{2\widetilde{\sigma}^4} \cong \chi^2_{(p-1)},$$

onde SQE é a soma de quadrados explicada na regressão auxiliar e $\widetilde{\sigma}^4$ é o quadrado de: $\widetilde{\sigma}^2 = \frac{\sum \varepsilon_i^2}{n}$. Um valor amostral da estatística de Breusch-Pagan maior do que o valor critico, para $p-1$ graus de liberdade e para o nível de significância pretendido, levará à rejeição da hipótese nula e à conclusão pela presença de heteroscedasticidade, [31, 5]. Isto é, caso se tenha que $BP_{obs} > \chi^2_{(1-\alpha,(p-1))}$, rejeita-se a hipótese de homoscedasticidade dos resíduos, considerando um nível de significância α.

Salienta-se que o teste de Breusch-Pagan tem apenas validade assintótica. Assim, a inexistência de indicações seguras e generalistas quanto à qualidade da aproximação em amostras de dimensão finita, implica que, nestes casos, por vezes, pode levar a conclusões contraditórias. Não obstante, no limite, o teste pode conduzir a conclusões idênticas sobre a existência de heteroscedasticidade, [45].

A violação do pressuposto de homoscedasticidade dos resíduos implica que os parâmetros estimados pelo método dos mínimos quadrados são não enviesados. Desta forma, sendo comprometida a distribuição dos resíduos do modelo, verifica-se que o cálculo de intervalos de confiança e testes de significância não é possível, [39].

Uma vez que o teste de Breusch-Pagan é sensível ao pressuposto da normalidade, Koenker (1981) propõe um teste, (**teste de Koenker**) baseado num estimador mais robusto.

Consideram-se os resíduos $\hat{\varepsilon}_i$, para $i = 1, ..., n$. Caso a hipótese nula seja verdadeira, isto é, caso exista homoscedasticidade nos resíduos, tem-se que:

$$\hat{\sigma^2} = \frac{1}{n} \sum \hat{\varepsilon}_i^2,$$

designa a estimativa da variância. Sejam $A = \frac{\hat{\varepsilon}_i^2 - \hat{\sigma^2}}{n}$ e $\bar{y} = \frac{1}{n} \sum \hat{y}_i$. Assim, a estatística de teste é dada por:

$$V = \frac{\left\{\sum \hat{\varepsilon}_i^2 (\hat{y}_i - \bar{y})\right\}^2}{A \sum (\hat{y}_i - \bar{y})^2}, \tag{3.37}$$

a qual segue, aproximadamente, uma distribuição qui-quadrado com um grau de liberdade quando a hipótese de homocedasticidade é verdadeira. Considerando um nível de significância α, rejeita-se a hipótese nula se $V_{obs} \geq \chi^2_{1-\alpha,1}$.

iv) ***Outliers* dos resíduos**

A partir da análise gráfica dos resíduos é também possível identificar eventuais *outliers* nos mesmos, os quais podem constituir observações influentes, sendo por vezes responsáveis pela invalidação dos pressupostos inerentes ao modelo de regressão linear múltipla.

Assim, *outliers* são observações extremas, não características, referentes a resíduos que são consideravelmente superiores (em valor absoluto) aos resíduos relativos às restantes observações. Os efeitos dos *outliers* podem ser variáveis, em função da sua localização, pelo que é plausível afirmar que o local onde se encontra o *outlier* determina a severidade da sua influência sobre a estimação dos coeficientes de regressão linear múltipla. Consequentemente, os efeitos dos *outliers* dizem-se moderados caso estes se encontrem no meio do domínio das observações; ou, por outro lado, podem ser extremos, caso se encontrem próximos dos limites do domínio das observações, [52, 39].

Nesse sentido, por vezes a invalidação dos pressupostos de regressão linear é devida à existência de observações influentes (*outliers*) na respetiva variável de resposta. Com efeito, a remoção dessas observações pode resultar na validação de um ou mais pressupostos do modelo de regressão linear, [16]. A deteção das observações influentes pode ser concretizada, numa primeira fase, através da representação gráfica dos resíduos através de gráficos de dispersão ou de caixas de bigodes; e posteriormente, pode ser usada uma medida diagnóstico, como é exemplo a **distância de Cook**. Esta permite avaliar a influência de uma observação x_{ij} sobre a estimação de β.

A **distância de Cook** consiste na alteração estandardizada dos valores ajustados. Assim, cada elemento na distância de Cook representa uma alteração normalizada no vetor de coeficientes, resultante da exclusão de uma observação, [12, 3]. Com efeito, a distância de Cook, DC_i, correspondente à observação i, é dada por:

$$DC_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}, \tag{3.38}$$

onde

- $\hat{y}_j$ representa o j-ésimo valor de resposta predito;
- $\hat{y}_{j(i)}$ é o j-ésimo valor predito, cujo cálculo não inclui a observação i;
- MSE é o erro quadrático médio;
- p corresponde ao número de coeficientes no modelo de regressão.

Algebricamente, a distância de Cook é equivalente à seguinte expressão:

$$DC_i = \frac{r_i^2}{pMSE} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right), \tag{3.39}$$

onde r_i é o i-ésimo resíduo e h_{ii} é o elemento da diagonal da matriz $(X^T X)^{-1} X$, isto é, $\hat{X}$, sendo que representa o valor de *leverage*[1]. Com efeito, ambos os valores x e y do conjunto de dados estão envolvidos no cálculo da distância de Cook, [33].

Assim, DC_i possui informação relativamente à influência da remoção da i-ésima observação do conjunto de dados, nos valores preditos. Desta forma, um ponto que possua um valor de DC_i elevado (superior a 1) indica que essa observação é extremamente influente na estimação dos coeficientes de regressão, [16, 12].

3.4.5.2 Ortogonalidade das variáveis independentes

Por vezes, algumas variáveis usadas no modelo de regressão linear múltipla construído podem ser exatamente, ou aproximadamente, combinações lineares das restantes variáveis explicativas, e, consequentemente, encontram-se fortemente correlacionadas entre si. A este fenómeno designa-se comummente por **colinearidade**, [47, 70].

Neste contexto, a hipótese inerente ao presente pressuposto é escrita por, [45]:

Na amostra de dimensão $n > K$, as variáveis explicativas $X_1, X_2, ...X_K$ e o termo independente (se incluído) são linearmente independentes.

Assim, em cada relação de colinearidade, existe uma variável redundante, o que torna a análise do modelo de regressão extremamente confusa e desprovida de significado. Desta forma, a sua validação encontra-se no cerne da análise e validação de um modelo de regressão linear múltipla, sendo que, idealmente, as variáveis explicativas do modelo de regressão não se encontram correlacionadas, isto é, são ortogonais, [52]. O uso dos métodos de seleção de preditores, abordados na secção 3.4.3, permitem evitar a existência de colinearidade das variáveis explicativas, [47, 70]. Não obstante, por vezes tal não é suficiente, sendo necessário detetar a existência deste fenómeno, o qual pode ser desenvolvido, por exemplo, recorrendo às seguintes metodologias:

- Numa primeira fase, a forma mais simples de averiguar a existência de colinearidade entre as variáveis explicativas baseia-se na construção e análise da matriz de correlações entre as variáveis explicativas, ou dos gráficos de dispersão. A ortogonalidade entre as variáveis explicativas é detetada pela existência de valores reduzidos de correlação e pela maior dispersão dos valores nos gráficos entre as variáveis explicativas, respetivamente.

- A existência de colinearidade entre as variáveis explicativas pode ser verificada recorrendo a metodologias estatísticas, mormente, através do **VIF** –*Variance Inflation Factor* (fator de inflação da variância). Este valor quantifica o efeito da correlação de uma variável com as restantes variáveis na inflação do erro padrão de um coeficiente de regressão, [20]. Assim, o valor de VIF é calculado pela expressão:

$$VIF = \left(\frac{1}{1 - R_i^2} \right),$$

onde R_i^2 representa o coeficiente de determinação múltiplo do modelo, entre X_i (considerada como variável resposta) e as restantes X_j variáveis, com $j \neq i$, $i = 1, \ldots, K$.

[1] O valor *leverage* é uma medida de quão distantes se encontram os valores das variáveis independentes correspondentes a uma observação, dos restantes valores correspondentes às restantes observações.

Na prática, valores de VIF superiores a 10 são representativos de problemas na estimação de β_i devido a problemas de colinearidade nas variáveis independentes, [26, 52].

Assim, para contornar a existência de colinearidade entre as variáveis explicativas do modelo de regressão, numa primeira fase, a construção do mesmo deve ser pautado por uma escolha cuidada e assertiva, baseada no conhecimento científico e na análise exploratória nas variáveis a usar no modelo. O cálculo e correspondente eliminação das variáveis cujos valores de VIF sejam superiores a 10, pode também minimizar esse efeito. Em casos extremos, por vezes apenas a aplicação de outras metodologias de regressão, como é exemplo a regressão de Ridge, permite colmatar esta problemática, [23, 47, 70].

A existência de ortogonalidade entre as variáveis explicativas de um modelo de regressão permite utilizar os coeficientes de regressão estimados pelo métodos dos mínimos quadrados podem utilizar-se com fins inferenciais e de estimação, [52, 9, 45].

A verificação dos pressupostos supracitados, permite efetuar uma interpretação do modelo de regressão de forma mais fidedigna, permitindo ainda averiguar a possibilidade do uso de resultados de inferência estatística com vista ao cálculo de intervalos de confiança e de predição (adequada caso os pressupostos sejam válidos).

3.4.6 Modelo de regressão linear com variáveis categóricas

Em inúmeros cenários de análise de regressão é necessário incluir variáveis qualitativas ou categóricas no conjunto de variáveis explicativas do modelo, para além das variáveis de índole quantitativa - comummente usadas nos referidos modelos - que influenciam a variável de resposta. A estimação dos coeficientes de regressão caso se verifique a inclusão de variáveis qualitativas segue uma lógica semelhante à referida estimação caso apenas sejam consideradas variáveis de índole quantitativa, [62, 72].

Assim, tipicamente num modelo de regressão consideram-se, concomitantemente, variáveis explicativas de índole quantitativa e qualitativa. Desta forma, o objetivo inerente a um modelo de regressão categórica consiste em analisar a influência de um grupo de uma variável categórica na variável de resposta Y, considerando as restantes variáveis (qualitativas e quantitativas) constantes, [20].

A dificuldade que surge nestes casos é a de que os fatores não podem ser representados por variáveis com domínio num conjunto contínuo de números reais, ao contrário do que se verifica com variáveis explicativas de índole quantitativa. Desta forma, a diferença entre modelos de regressão que incluem variáveis explicativas *dummy*, e modelos que apenas incluem variáveis explicativas de cariz quantitativo, reside na necessidade de codificação das variáveis categóricas, sendo que a existência de múltiplos níveis inerentes às mesmas implica a definição de um grupo (nível) de referência, [39].

Por questões de simplicidade, suponha-se que, na construção de um modelo de regressão, tem-se uma variável explicativa categórica Z, a qual possui três categorias Z_1, Z_2 e Z_3, e ainda uma variável quantitativa, X_j, com $j = 1, ..., K$. Considerando Z_3 o grupo de referência inerente à variável Z, e sendo Z_1 e Z_2 variáveis *dummy* referentes às categorias Z_1 e Z_2, respetivamente,

o modelo de regressão pode ser escrito como:

$$E(Y) = \beta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \beta_1 X_1 + ... + \beta_K X_K \tag{3.40}$$

Com o intuito de abordar a interpretação dos coeficientes e de visualizar o modelo de regressão, seguidamente analisa-se a equação (3.40) considerando apenas uma variável quantitativa, X, pelo que o modelo é dado por:

$$E(Y) = \beta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \beta_1 X. \tag{3.41}$$

Primeiramente apresenta-se o modelo de regressão passível de ser escrito para cada grupo em separado, através da substituição dos valores de cada variável *dummy* na equação (3.41). Assim, a equação relativa ao grupo Z_3 vem:

$$E(Y) = \beta_0 + \delta_1(0) + \delta_2(0) + \beta_1 X = \beta_0 + \beta_1 X. \tag{3.42}$$

Para o grupo Z_1, tem-se

$$E(Y) = \beta_0 + \delta_1(1) + \delta_2(0) + \beta_1 X = \beta_0 + \delta_1 + \beta_1 X. \tag{3.43}$$

E, por fim, para o grupo Z_2, a equação de regressão é dada por:

$$E(Y) = \beta_0 + \delta_1(0) + \delta_2(1) + \beta_1 X = \beta_0 + \delta_2 + \beta_1 X. \tag{3.44}$$

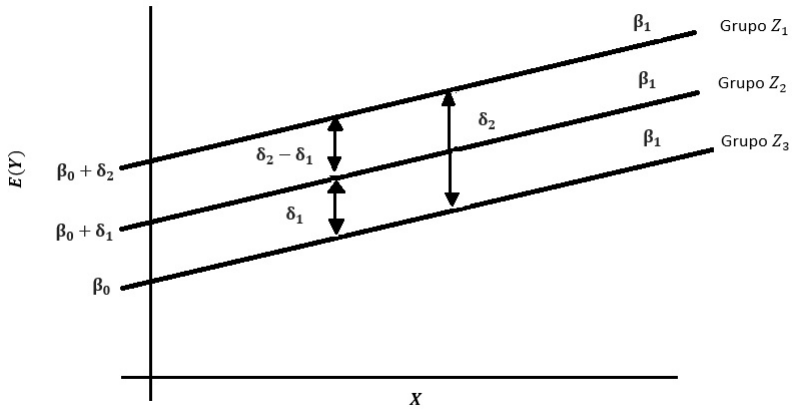


Figura 3.2: Modelo de regressão $E(Y) = \beta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \beta_1 X$.

A figura 3.2 permite intuir quanto à interpretação deste modelo. Assim, as três equações constituem, essencialmente, modelos de regressão linear simples, com Y a variável dependente e X a variável independente, [20]. É de realçar que cada equação referente ao respetivo nível do grupo da variável categórica possuem ordenada na origem distinta, porém apresentam o mesmo declive, β_1, o que implica que as retas sejam paralelas (visível no gráfico da figura 3.2).

Desta forma, a reta de regressão entre Y e X tem ordenada na origem β_0 no grupo Z_3, $\beta_0 + \delta_1$ no grupo Z_1 e $\beta_0 + \delta_2$ no grupo Z_2. Ambas as retas possuem declive β_1. Consequentemente, o efeito de X é o mesmo em cada grupo da variável Z. Por outro lado, δ_1, δ_2 e $\delta_2 - \delta_1$ representam diferenças constantes na média de Y entre os grupos, independentemente do valor de X. Com efeito, a diferença da média de Y entre os grupos Z_1 e Z_3 é dada por:

$$E(Y|Z_1) - E(Y|Z_3) = \beta_0 + \delta_1 + \beta_1 X - (\beta_0 + \beta_1 X) = \delta_1.$$

E para os grupos Z_2 e Z_3 tem-se

$$E(Y|Z_2) - E(Y|Z_3) = \beta_0 + \delta_2 + \beta_1 X - (\beta_0 + \beta_1 X) = \delta_2.$$

E, por fim, para os grupos Z_1 e Z_2 tem-se

$$E(Y|Z_2) - E(Y|Z_1) = \beta_0 + \delta_2 + \beta_1 X - (\beta_0 + \delta_1 + \beta_1 X) = \delta_2 - \delta_1.$$

A título conclusivo é plausível afirmar que os deltas e as suas diferenças, representam as alterações na média de Y entre os grupos, após o ajustamento da variável de referência, [39, 62]. Na figura 3.2 essas diferenças encontram-se representadas pelas duas setas. Note-se que estas diferenças entre os grupos na média de Y, $E(Y)$, são as mesmas para vários valores de X. Tal representa uma característica inerente ao modelo de regressão com variáveis *dummy*, visto que este não permite considerar a interação entre X e Z.

Salienta-se por fim que a exposição teórica apresentada faz uso do modelo aditivo de regressão com variáveis categóricas, visto que, a abordagem prática faz uso de tais metodologias. Seria ainda possível considerar o modelo de regressão multiplicativo ou misto, tendo por base a relação das variáveis na equação de regressão. Com efeito, num modelo de regressão multiplicativo, a equação de regressão é escrita por $E(Y) = \beta_0 + \delta_1 Z_i X + \beta_1 X$, e em termos geométricos as retas de regressão intercetam-se, o que postula que a diferença entre as variáveis categóricas à medida que se verificam variações na variável quantitativa tende a variar. O modelo de regressão com variáveis *dummy* na forma combinada aditiva e multiplicativa permite garantir maior flexibilidade do modelo.

3.5 Árvores de regressão

Em inúmeras situações, a complexidade inerente aos conjuntos de dados, a existência de desequilíbrio e de valores ausentes, pode implicar a impossibilidade de explicar a relação existente entre as variáveis de forma linear, [8]. Para além da não linearidade da relação entre as variáveis, pode ainda ocorrer que esta envolva interações muito elevadas, o que constitui um impedimento à aplicação de metodologias de análise de regressão linear.

Assim, as árvores de regressão são técnicas de análise estatística mais recentes, idealmente usadas na exploração e modelação de dados com as características explanadas, permitindo compreender as relações estruturais entre a variável resposta e as variáveis preditivas, e, ainda, prever a variável resposta da forma mais exata possível, [13, 58].

3.5.1 Conceitos introdutórios

Os modelos de classificação e regressão em árvores oferecem uma alternativa para estudar a relação entre uma variável dependente e uma ou mais variáveis preditivas, caso esta não possa ser estudada recorrendo a modelos lineares. A diferença entre árvores de regressão e árvores de classificação, reside na índole da variável dependente. Assim, apesar de ambas as árvores de regressão e classificação assumirem variáveis explicativas de cariz qualitativo ou quantitativo, no que concerne à variável dependente, em árvores de regressão é de índole quantitativa, enquanto que em árvores de classificação é de cariz qualitativo, [18, 19].

Uma árvore de regressão faz uso da estratégia *dividir para conquistar* na sua construção, pelo que um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada uma estratégia análoga, e assim sucessivamente. Desta forma, as soluções dos subproblemas podem ser combinadas, na forma de uma árvore, para produzir uma solução do problema original, [25]. Assim, existem algumas técnicas de classificação e regressão em árvore, que diferem quanto ao método usado para a divisão do conjunto de dados, das quais se destacam: ID3 (Quinlan, 1979), ASSISTANT (Cestnik *et al.* 1987), CART (*Classification and Regression Trees*, Breiman *et al.* 1984) e a CHAID (*Chi-Square Automatic Iterative Detection*).

Na presente investigação ir-se-á dar maior ênfase à técnica CART, na qual os dados são divididos repetidamente e sequencialmente, de forma a que os subgrupos resultantes de cada divisão apresentem entre si a maior heterogeneidade possível e a maior homogeneidade interna. Consequentemente, a árvore é construída repetindo sucessivas divisões dos dados, através de uma regra simples baseada numa única variável preditiva. Em cada divisão os dados são separados em dois grupos mutuamente exclusivos, o mais homogéneos possível. O procedimento de divisão é repetido e aplicado separadamente a cada um dos grupos originados, [25, 19, 43].

Formalmente, uma árvore de regressão é um gráfico acíclico direcionado, composto por um nó raiz, um conjunto de nós interiores, e nós terminais, usualmente denominados nós folha. O nó raiz e os nós interiores, referidos coletivamente como nós não terminais, estão ligados em fases de decisão. Os nós terminais representam as classes finais. Assim, a árvore de regressão representa um conjunto de restrições ou condições que são hierarquicamente organizadas, e que são aplicadas sucessivamente a partir de uma raiz a um nó terminal ou folha da árvore, [4].

O objetivo inerente à presente metodologia consiste na construção de partições da variável resposta em grupos homogéneos, mantendo a árvore razoavelmente pequena. A divisão é

desenvolvida repetidamente até ser obtida uma árvore de grande dimensão, com inúmeros grupos finais, a qual é posteriormente podada até que seja obtida uma árvore ótima com a dimensão desejada, [6, 25].

A formação dos grupos a partir das variáveis preditivas depende do seu tipo. Assim, no caso de variáveis explicativas categóricas com 2 níveis, apenas é possível uma divisão. Caso a variável explicativa possua $k > 2$ níveis, existem $2^{k-1} - 1$ possibilidades de divisão. Por outro lado, caso as variáveis explicativas sejam quantitativas, a divisão baseia-se na verificação de uma condição do tipo $x_i > v$, ou $x_i < v$, sendo que para este valor v são considerados todos os valores intermédios da correspondente variável preditiva. Note-se que, de entre todas as possíveis divisões considerando todas as variáveis explicativas do modelo, seleciona-se a que torna máxima a homogeneidade dos grupos resultantes. Esta é definida e avaliada de várias formas, e a sua escolha depende do tipo de variável resposta, [18, 8].

Nas subsecções seguintes, ir-se-á considerar Y como a variável dependente, de índole numérica, e as variáveis explicativas, X_i, com $i = 1, ..., k$, onde k designa o número de preditores.

3.5.2 Construção de árvores de regressão

A árvore é construída a partir de sucessivas divisões dos dados, definida por uma regra específica sob uma variável explicativa. Em cada divisão os dados são particionados em dois grupos mutuamente exclusivos, cada um dos quais é tão homogéneo quanto possível. O processo de divisão é então aplicado a cada grupo separadamente. O objetivo consiste em particionar a variável resposta em grupos homogéneos, e, concomitantemente, manter a árvore razoavelmente pequena. O tamanho da árvore traduz o número de grupos finais. A divisão prossegue até que seja construída uma árvore suficientemente grande, a qual é depois podada até ao tamanho desejado, [13].

Assim, o algoritmo inerente à construção das árvores de regressão apresenta-se seguidamente.

Algoritmo 1: Pseudocódigo usado na construção da árvore de regressão (técnica CART).

1 Iniciar no nó raiz.
2 Para cada X_i, procurar o conjunto S de variáveis que minimiza a soma das impurezas dos nós, nos dois nós recém formados, e escolher a divisão X^* ($X^* \in S$) que minimiza globalmente X e S.
3 Se o critério de paragem for satisfeito, a divisão termina. Caso contrário, voltar ao passo 2 para cada um dos nós recém formados.

Assim, na construção de uma árvore de regressão devem ser considerados os seguintes aspetos:

1. Regra de divisão para a regressão, isto é, seleção do ”melhor”critério para a divisão;
2. Critério de paragem;
3. Atribuição de um valor aos nós folha.

Estes aspetos serão abordados nas secções subsequentes.

Regra de divisão para a regressão

A homogeneidade dos nós é definida por uma medida de impureza, a qual assume o valor zero em nós completamente homogéneos e aumenta à medida que a homogeneidade diminui. Existem inúmeras medidas de impureza definidas na literatura, sendo que estas variam caso se trate de um modelo de classificação ou regressão. Assim, uma função custo é frequentemente associada a uma regra de divisão da árvore, [58].

Em problemas de regressão, são propostos dois critérios para a seleção da melhor regra de divisão, com base nas estimativas dos erros obtidos. Com efeito, a função de custo a minimizar é, usualmente, o erro quadrático médio (EQM) e o erro absoluto médio (EAM) também designado por desvio absoluto médio. Estes dois critérios designam medidas de impureza na medida em que assumem valores nulos para nós completamente homogéneos, e aumentam à medida que a homogeneidade dos grupos diminui, [43, 8]. Neste contexto, é premente minimizar o EQM ou EAM, por forma a maximizar a homogeneidade. Supondo que n designa a dimensão dos dados em estudo, seguidamente analisam-se as medidas de impureza supracitadas.

- **Erro Quadrático Médio (EQM)**: este critério é semelhante à minimização dos mínimos quadrados no modelo de regressão linear. Assim, as divisões são escolhidas por forma a minimizar a soma dos quadrados dos erros entre as observações reais e a média de cada nó, o que se traduz na minimização da equação subsequente.

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{3.45}$$

  onde $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

- **Erro Absoluto Médio (EAM)**: este método minimiza o desvio médio absoluto da mediana em cada nó, o que se traduz na minimização da equação seguinte:

$$EAM = \sum_{i=1}^{n} |Y_i - m_e|, \tag{3.46}$$

  onde m_e designa a mediana das n observações Y_i.

  A vantagem deste método em detrimento do EQM, reside no facto de não ser tão sensível à presença de *outliers* e permitir a obtenção de um modelo mais robusto. Porém, a desvantagem associada a este método reside no facto de ser pouco eficiente quando os dados possuem uma proporção elevada de valores nulos e as variáveis explicativas forem categóricas, [56].

Assim, procura-se o teste/regra, experimentando todas as variáveis preditivas e tomando todos os valores intermédios para v, da condição $X_i > v$ ou $X_i \leq v$, correspondentes a essas variáveis que minimizam o EQM ou EAM.

Critério de paragem

O erro associado ao teste t pode ser definido da seguinte forma, supondo n elementos:

$$Erro(n_t) = \frac{1}{n} \sum_{i \in P_n} (Y_i - \bar{Y}_n)^2, \tag{3.47}$$

onde $\bar{Y}_n$ designa o valor médio da variável dependente, Y, calculada no conjunto P_n da partição considerada à regra t.

A aplicação do critério de paragem pode ser implementada segundo dois procedimentos:

- A divisão prossegue até que o "ganho"devido à nova divisão seja inferior a um valor pré-definido. Este procedimento possui a desvantagem de que, caso a regra de paragem seja baseada num número reduzido, esta metodologia conduz à construção de árvores de dimensões demasiado elevadas, o que implica *overfitting* do modelo. Por outro lado, caso o número escolhido seja muito elevado, a árvore construída pode não expressar devidamente as interações entre as variáveis explicativas, [18].

- Segundo este procedimento, a construção da árvore para se, numa divisão, o erro existente antes da divisão for inferior ao erro obtido após a divisão. Assim, nesta metodologia, sendo o erro antes da divisão definido por

  $$Erro_{antes} = Erro(n_t),$$

  e o erro após a divisão definido por

  $$Erro_{apos} = \frac{n_{t,e}}{n_t} Erro(n_{t,e}) + \frac{n_{t,d}}{n_t} Erro(n_{t,d}).$$

  Em cada divisão comparam-se os erros correspondentes, e, caso $Erro_{apos} > Erro_{antes}$, a construção da árvore termina, [6].

3.5.3 Validação das árvores de regressão

O processo de subdivisão sucessiva da amostra implica que, em cada iteração, o número de dados seja cada vez mais reduzido. Neste contexto, dois problemas inerentes à árvore construída podem ser destacados: a falta de fiabilidade das estimativas obtidas para os erros, e o problema de *overfitting*. Assim, com o intuito de tentar resolver as problemáticas supracitadas, podem ser aplicadas as seguintes estratégias:

a) Uso de estimativas mais consistentes para o erro;

b) Inclusão de métodos de avaliação mais fiáveis, no critério de paragem do crescimento das árvores;

c) Construção de uma árvore de regressão de elevada dimensão, e posteriormente podar, fazendo uso de um critério adequado para tal.

Consistência do erro

Objetivando-se contornar a falta de consistência inerente às estimativas do erro, devido à diminuição da dimensionalidade do conjunto de dados à medida que o procedimento de construção da árvore prossegue, propõem-se seguidamente métodos que aumentam a consistência dos erros e a fiabilidade da árvore de regressão construída.

i) **Método de Re-amostragem** (o método de *Holdout*):

O uso da presente metodologia é indicado quando o conjunto de dados possui dimensões elevadas, visto que caso a dimensão da amostra seja reduzida, o erro calculado na predição pode sofrer uma grande variação.

Esta metodologia é frequentemente usada, e consiste na divisão da amostra em dois subconjuntos, mutuamente exclusivos, um conjunto de treino usado para estimação, ou seja, para construir as árvores, e o conjunto de teste para obter as estimativas do erro, isto é, para testar o modelo. Habitualmente, o conjunto de treino contém 60% das observações, e o conjunto de teste contém as restantes observações do conjunto de dados, sendo que a construção dos dois conjuntos faz uso de amostragem aleatória.

A vantagem inerente a esta metodologia consiste no uso de amostras independentes na construção de árvores e na estimação dos erros. Não obstante, salienta-se que este procedimento pode também constituir uma desvantagem na medida em que a obtenção da árvore de regressão faz uso de menor quantitade de dados, [4, 6].

ii) **Validação Cruzada**

Tendo em conta que o número de observações disponíveis nem sempre é suficiente para a constituição de amostras de treino e teste, o uso de uma validação cruzada surge como uma alternativa de validação do modelo quando a dimensão da amostra é reduzida. Nesse sentido, a validação cruzada torna possível o cálculo de um erro mais realista para a árvore construída.

A presente metodologia consiste na divisão do conjunto de dados em p subconjuntos com a mesma dimensão. São construídas p árvores diferentes, sendo que, na construção de cada uma, usam-se $p-1$ conjuntos para estimação e um para validação, sendo que este procedimento se repete p vezes. A partir da construção das p árvores, é possível calcular o erro associado a todas as observações, sendo que a estimativa obtida pra os erros é a média das p estimativas parciais. Geralmente considera-se $p = 10$.

Apesar de ser computacionalmente "pesado", sendo o método cujas estimativas são mais fiáveis, é o mais utilizado.

iii) ***Leave-one-out***

Designa um caso particular da validação cruzada, onde o número de subconjuntos coincide com o número de elementos da amostra. É usado preferencialmente quando o conjunto de dados apresenta dimensões reduzidas.

Critérios de paragem mais eficazes

Com o intuito de usar critérios de paragem mais eficazes, deve-se melhorar as estimativas do erro usadas. Assim, para o cálculo do erro (medido pelo EQM) pode-se usar, por exemplo,

o método de re-amostragem. Um procedimento análogo pode também ser implementado para calcular o erro após a divisão, $Erro_{apos}$, o que consiste na substituição dos erros anteriormente especificados pelos respetivos erros obtidos através do método de re-amostragem ou de validação cruzada. Neste contexto, o crescimento da árvore deve parar se, numa determinada divisão, a condição $Erro_{apos} \geqslant Erro_{antes}$ for verdadeira, o que implica que, nestas circunstâncias, a divisão do nó não seria proveitosa no que concerne ao erro, [25].

Estratégias de poda de uma árvore de regressão

A poda é considerada a fase mais importante do processo de construção da árvore. Assim, esta metodologia apresenta-se como central na prevenção do problema de *overfitting* e no melhoramento das previsões. Com efeito, podar uma árvore, isto é, trocar nós profundos por folhas, ajuda a minimizar os problemas supracitados, [25].

A poda das árvores de regressão consiste num procedimento *"standard"* neste tipo de metodologias cujo objetivo inerente consiste em proporcionar um melhor compromisso entre a simplicidade e compreensibilidade das árvores e a sua capacidade preditiva. Com efeito, o desenvolvimento de metodologias de poda permite a obtenção de melhores resultados no que concerne ao erro do modelo. Permite também a "inspeção"de um conjunto de modelos alternativos que diferem relativamente ao compromisso tamanho/erro, [8].

Os métodos de poda podem ser divididos em dois grupos principais:

- **Pré-poda:**

  Este grupo inclui os métodos que param a construção da árvore quando algum critério é satisfeito, pelo que, a pré-poda faz uso de regras de paragem que previnem a construção daqueles ramos que não parecem melhorar a precisão preditiva da árvore. A pré-poda tem vantagem de não perder tempo na construção de uma estrutura que não é usada na árvore final. Desta forma, fazer a poda da árvore consiste em: obter uma árvore exageradamente grande; gerar uma sequência de sub-árvores; e, por fim, escolher a "melhor"sub-árvore, usando métodos de avaliação fidedignos na escolha do modelo final.

- **Pós-poda:**

  Este é o método mais comum de poda de árvores de regressão, sendo composto por técnicas que constroem uma árvore completa e, posteriormente, realizam a respetiva poda. Consequentemente, num método pós-poda, cortam-se os ramos da árvore desnecessários após esta estar completa. Apesar de ser o método mais usado em inúmeros sistemas e pacotes informáticos, em problemas de elevada dimensão é computacionalmente pouco eficiente.

  De entre as metodologias de pós-poda existentes, salienta-se a poda **custo - complexidade**, a qual foi apresentada por Breiman *et al.* (1984), sendo um dos métodos mais utilizados. Assim, inicialmente é gerada uma árvore completa. Com base nessa árvore é gerada uma sequência de árvores cada vez menores, sendo escolhida uma das subárvores. Este método baseia-se em dois parâmetros: a taxa de erro $R(T)$ e o tamanho da árvore, $|T|$, medido em termos das folhas. A medida de custo-complexidade para a árvore é:

$$R_\alpha(T) = R(T) + \alpha|T|, \tag{3.48}$$

onde α é um parâmetro que mede a importância relativa do tamanho da árvore em relação à taxa de erro. Assim, para cada valor de α, o objetivo é encontrar a subárvore $|T_\alpha| \leq |T|$ que minimiza $R_\alpha(T)$. Caso α assuma valores reduzidos, a penalidade é menor (visto que possui um número elevado de nós terminais), enquanto que T_α assume valores elevados. À medida que α aumenta, a subárvore minimizada será formada por um número reduzido de nós terminais. Ainda que α assuma valores contínuas, tem-se que o número de subárvores de T é finito. O processo de poda produz, assim, uma sequência finita de subárvores de T, $T_1, T_2, ...$ com nós terminais progressivamente menores. Desta forma, valores elevados de α conduzem a uma árvore pequena, enquanto que valores reduzidos de α implicam uma árvore com elevada complexidade. A árvore podada que é selecionada é aquela que minimiza $R_\alpha(T)$ na sequência das subárvores construídas, [8, 25].

Recentemente diversos programas estatísticos, como é exemplo o *software* R, incorporaram funções que implementam árvores de decisão para problemas de classificação e regressão. Neste *software* os pacotes que incorporam funções que se encontram na génese de árvores de regressão são: 'rpart', '*tree*', 'Rweka', '*ipred*', 'evtree', entre outros. Na presente investigação fez-se uso do pacote 'rpart' do *software* R, a qual se fundamenta numa técnica robusta, baseada num algoritmo de custo-complexidade da árvore, explorado anteriormente. Assim, tendo por base a nomenclatura adotada por este pacote, o principal objetivo do algoritmo consiste na minimização da complexidade da árvore. Com efeito, a divisão é implementada procurando o valor máximo do parâmetro α que verifica

$$EC_\alpha(T) < xerror + xstd,$$

onde, $xerror$ representa o valor mínimo do erro da validação cruzada e $xstd$ o correspondente desvio padrão.

Assim, a metodologia implementada pelas funções do pacote 'rpart' do *software* R, determinam o valor do parâmetro de complexidade, o qual deve ser suficientemente pequeno por forma a que o erro de validação cruzada seja também mínimo. Após a identificação da árvore ideal, isto é, a árvore que possui erro de validação cruzada mínimo, é possível ainda podar o número de divisões.

Capítulo 4

O problema e os dados

"O trabalho do estatístico é catalisar o processo da aprendizagem científica"

George Box

No presente capítulo apresenta-se uma breve descrição do problema proposto pela entidade de acolhimento, RAIZ e do conjunto de dados a analisar. É ainda efetuada uma breve descrição das variáveis usadas no desenvolvimento das análises subsequentes e do processamento implementado no conjunto de dados. Por fim analisam-se, sumariamente, as áreas de Portugal Continental em estudo na presente investigação.

4.1 Formulação do problema em estudo

O presente estudo objetiva evidenciar os benefícios económicos da gestão florestal em áreas de minifúndio. Com esse intuito, pretende-se avaliar o efeito de níveis diferenciados de gestão florestal nos indicadores de produtividade. Assim, os estudos desenvolvidos pretendem verificar de que forma a existência de gestão florestal efetuada por diferentes entidades, influencia a produtividade florestal, e consequentemente, maximizar os benefícios económicos inerentes à exploração florestal.

Para o presente estudo utilizaram-se conjuntos de dados que colecionam informação relativamente ao inventário florestal de parcelas pertencentes a diferentes entidades cuja localização permite identificar duas áreas de estudo. A descrição dos conjuntos de dados é implementada na secção subsequente.

4.2 Descrição dos conjuntos de dados

Os conjuntos de dados fornecidos agregam informação relativa a dados de inventário florestal, ou seja, informações relativas a inúmeras características biométricas das árvores das parcelas inventariadas. Com efeito, cada observação dos conjuntos de dados fornecidos faz referência a uma parcela inventariada, agregando informação relativa às árvores aí existentes.

Os conjuntos de dados fornecidos são constituídos por informações de inventário florestal de parcelas pertencentes a empresas que fazem gestão florestal e organizações de produtores florestais de uma região do centro de Portugal. Tendo por base o cariz de confidencialidade inerente à presente investigação, as referidas entidades responsáveis pelas parcelas são designadas por entidade A, B e C, de forma aleatória.

4.2.1 Conjunto de dados 'Dados Inventário 2017'

O primeiro conjunto de dados, cujo inventário florestal teve lugar em novembro de 2017, é formado por 233 parcelas, as quais pertencem a duas entidades distintas, a entidade A e a entidade B. Nas 233 parcelas verifica-se que a aplicação de técnicas de gestão florestal não é comummente aplicada pelas duas entidades, e, por conseguinte, tem-se que, em algumas parcelas foram aplicadas técnicas de gestão florestal, enquanto que noutras parcelas, existem evidências da não aplicação de técnicas de gestão florestal. A classificação das parcelas inventariadas como geridas ou não geridas é realizada com base na existência de elementos indicativos à data do inventário (existência de vegetação alta, falta de controlo do mato, entre outros), e conjugada com informação existente na base de dados.

Este conjunto de dados é formado por 60 variáveis, através das quais é possível descrever inúmeras características das parcelas inventariadas. O desígnio atribuído ao presente conjunto de dados é '**Dados Inventário 2017**'.

4.2.2 Conjunto de dados 'Dados entidade C'

O segundo conjunto de dados é formado por 377 parcelas, pertencentes, exclusivamente, a uma entidade, designada por entidade C. A totalidade das suas parcelas pauta-se pela existência de evidências da aplicação de técnicas de gestão florestal. O referido conjunto de dados possui informações sobre o inventário florestal desenvolvido entre os anos de 2002 e 2015. É formado por 38 variáveis que permitem descrever as parcelas inventariadas. Este conjunto de dados é designado por '**Dados entidade C**'.

4.2.3 União dos conjuntos de dados

Uma análise crítica, baseada em gráficos e conhecimentos a nível florestal, tornou possível a deteção da existência de valores incongruentes nos conjuntos de dados. Assim, tendo sido detetadas algumas parcelas cujos valores recolhidos não se coadunam com o que seria expectável, na medida em que globalmente seria impossível serem verificadas tais situações, ou caso os valores calculados relativamente a algumas variáveis fossem omissos, as respetivas parcelas foram removidas. Com efeito, removeram-se parcelas com idades reduzidas (visto que não permitem adicionar informação pertinente ao estudo) e parcelas cujo valor de AMAutil12 se encontrava omisso (tendo em conta que é uma variável preponderante na presente investigação). Assim, no total foram eliminadas 67 parcelas do conjunto de dados '**Dados Inventário 2017**', pelo que, em detrimento das 233 parcelas inventariadas originalmente existentes no referido conjunto de dados, foram usadas 166 parcelas nos estudos que se seguem. Relativamente ao conjunto de dados '**Dados entidade C**', também foram detetados alguns valores discrepantes (*outliers*) no que infere às variáveis AMAutil12, Vu12 e S a partir da observação das caixas de

bigodes que se encontram no Anexo C.1. Procedeu-se à sua identificação e foram implementadas as mesmas análises com e sem remoção dos referidos valores, donde se concluiu que os resultados obtidos após a remoção dos *outliers* se coadunam com os resultados obtidos sem proceder à sua remoção. Salienta-se que os *outliers* detetados correspondem a parcelas cujos valores de AMAutil12 são mais elevados. Desta forma, não se considerou preponderante proceder à eliminação dos *outliers* no conjunto de dados '**Dados entidade C**' visto que tal implica perda de informação.

Por forma a viabilizar as análises subsequentes, começou por se agregar os dois conjuntos de dados, '**Dados Inventário 2017**' e '**Dados entidade C**', após a remoção das parcelas cujos valores continham erros. Desse procedimento resultou um só conjunto de dados, o qual agrega informação sobre as parcelas pertencentes às três entidades, A, B e C. O conjunto de dados obtido é formado por 543 parcelas e 68 variáveis.

4.3 Descrição das variáveis

O conjunto de dados obtido é formado por 68 variáveis, as quais objetivam caracterizar a tipologia da gestão aplicada, bem como verificar a produtividade das parcelas, e ainda apresentar algumas observações relevantes.

Na tabela 4.1 é apresentada, por ordem alfabética, uma descrição das variáveis usadas consideradas preponderantes na presente investigação. Relativamente ao tipo de medição (mencionado na secção 2.3), tem-se que [b] designa medição Biométrica e [p] traduz medição obtida por processamento.

Tabela 4.1: Descrição das variáveis dos conjuntos de dados.

Nome da variável	Descrição	Unidade de medida
AE [b]	Desígnio da área de estudo à qual a parcela pertence. Assume dois valores: 1 (área de estudo 1) e 2 (área de estudo 2).	
AMAutil [p]	Acréscimo médio anual útil das árvores da espécie *Eucalyptus Globulus* existentes na parcela.	m^3/ha ano
AMAutil12 [p]	Acréscimo médio anual útil projetado aos 12 anos da parcela. Medida de referência usada na comparação de parcelas relativamente à sua produtividade.	m^3/ha ano
Ano [b]	Ano da realização do inventario florestal da parcela.	
Controlo_Vegetação [b]	Especifica a existência ou inexistência de controlo da vegetação na parcela, pelo que assume dois valores possíveis: 'Com Controlo' e 'Sem Controlo'.	

Clima [b]	Classificação do tipo de clima existente na parcela em questão, sendo que varia entre 8 e 10. Estabelece-se que 8 classifica um clima que favorece menos a produtividade do que um clima classificado por 10 (melhor classificação de clima).	
Entidade [b]	Especifica a entidade responsável pela parcela, sendo que resulta do processamento final dos dados, através do cruzamento da informação recolhida pelos prestadores de serviços, relativa às entidades responsáveis pelas parcelas e por informações de registos. Assim, assume os valores A, B ou C.	
G12 [p]	Área basal das árvores dominantes da parcela em questão projetada aos 12 anos, ou seja, é a média das áreas basais das 100 árvores mais grossas por ha, projetada aos 12 anos, na parcela em questão.	m^2
gestão [b]	Resulta de um processamento final da informação recolhida pelos prestadores de serviços, de imagens de satélite, de informações de registos, e de fotografias das parcelas. Estabelece a inexistência ou existência de gestão de parcelas, assumindo para tal os valores 0 e 1, respetivamente.	
Hdom12 [p]	Altura dominante das árvores da espécie *Eucalyptus Globulus* existentes na parcela em questão, projetada aos 12 anos da referida espécie de eucalipto, consistindo assim numa medida de comparação entre diferentes parcelas.	m
Idade validada [p]	Idade da parcela, validada recorrendo à idade medida em campo, a imagens de satélite, conhecimentos da existência de fenómenos como incêndios, corte das árvores das parcelas, entre outros.	Anos
N [b]	Especifica o número de árvores vivas na parcela, sendo que, em caso da parcela se encontrar em talhadia, designa o número de varas vivas na parcela em questão.	
NPL [b]	Número de árvores plantadas na parcela em questão.	
Parcela [b]	Identificação do número da parcela associado à parcela em questão. A sua identificação é estabelecida aquando do lançamento da amostragem pelo instituto de investigação RAIZ.	

Preparação do Terreno [b]	Especificação do tipo de preparação do terreno aplicado na parcela em questão.	
Rotação [b]	Especificação do tipo de rotação inerente às árvores das parcelas em questão. São destacados dois tipos de rotação aplicados nas parcelas: primeira e talhadia.	
RP [b]	Designa a Região de produtividade da parcela, sendo que agrega a informação do clima, do solo e dos dias de precipitação, permitindo desta forma classificar a parcela em função das condições climatéricas e do solo existentes. Assume valores 2, 3 e 4, onde 2 classifica uma RP de excelência e 4 uma região de produtividade com qualidade inferior.	
S [p]	Site Index, ou seja, índice de qualidade da estação. É a altura de uma árvore "livre para crescer" de uma determinada espécie numa idade base na parcela em questão. Depende das espécies em questão, sendo o seu cálculo diferente em função das espécies existentes na parcela, pois verifica-se que entre espécies diferentes o Site Index não é equivalente.	m
Seleção de varas [b]	Especifica a existência ou inexistência de seleção de varas nas árvores da parcela em questão, assumindo os valores 'sim' e 'não', respetivamente.	
Solo [b]	Classifica a tipologia de solo existente na parcela em questão. É uma variável qualitativa/fator, e varia entre 5 e 7, onde 5 classifica um solo mais favorável do que 7.	
V12 [p]	Volume das árvores da espécie *Eucalyptus Globulus* com casca e com cepo projetado aos 12 anos da parcela, ou seja, designa o volume total das árvores da espécie *Eucalyptus Globulus* existentes na parcela em questão quando esta atingir os 12 anos.	m^3
Vmudi12 [p]	Volume mercantil, sem casca (sem cepo) das árvores da espécie *Eucalyptus Globulus* existentes na parcela em questão, para um diâmetro de desponta de di cm, projetado para os 12 anos da parcela em questão.	m^3
Vu12 [p]	Volume das árvores da espécie *Eucalyptus Globulus* sem casca e com cepo, existentes na parcela, projetado aos 12 anos.	m^3

Tendo por base a descrição das variáveis apresentada, destacam-se as variáveis N e NPL, que especificam o número de árvores vivas e o número de árvores plantadas nas parcelas, respetivamente. Desta forma, um olhar crítico sob as variáveis permite intuir que, a aplicação de metodologias de gestão florestal adequadas favorece o número de árvores sobreviventes na unidade parcelar. Assim, seria interessante, como índole de qualidade da parcela, analisar a quantidade de árvores sobreviventes numa parcela e comparar com o respetivo número em outras parcelas. Não obstante, o número de árvores plantadas em cada parcela varia, e, por conseguinte, o número de árvores vivas varia também, pelo que não é correto comparar esses valores entre parcelas diferentes. No sentido de colmatar essa problemática, foi criada uma nova variável, designada por FR_{nvivas}, que designa a frequência relativa do número de árvores que sobreviveram em cada parcela, a qual é obtida a partir da seguinte fórmula:

$$FR_{nvivas} = \frac{N}{NPL}$$

O valor obtido a partir do quociente explanado é um valor relativo, pelo que a variável criada permite efetuar a comparação da quantidade de árvores sobreviventes entre parcelas, constituindo desta forma um índice de qualidade da parcela. Por outras palavras, a variável construída designa a proporção de árvores que sobreviveram numa parcela. Uma análise breve da variável supracitada pode ser encontrada no Anexo C.2.

No Anexo C.3 apresenta-se uma descrição sumária das variáveis quantitativas e qualitativas do conjunto de dados. É ainda apresentada uma breve análise implementada sob a variável S, a pedido de investigadores do RAIZ, a partir da qual é plausível intuir que os valores médios de *Site Index* são incrementados em parcelas que se encontrem em primeira rotação, e onde existam evidências da existência de gestão florestal. As parcelas nas condições supracitadas que se encontrem em regiões de produtividade classificadas por 3 permitem incrementar o valor médio de *Site Index*, cujo valor inerente a esta variável consiste num valor de referência da produtividade das parcelas em condições de clima, temperatura e solo ideais.

4.4 Os grupos de gestão

Numa fase precedente à análise preliminar dos dados, com o intuito de dar resposta ao objetivo da presente investigação, e após a eliminação dos valores errados detetados, procedeu-se à classificação do conjunto de dados em cinco grupos, com base no cruzamento da informação relativa à gestão das parcelas e entidades responsáveis pelas mesmas.

No que concerne à informação relativa à gestão das parcelas, sendo a inteleção do referido conceito primordial na presente investigação, encontrando-se inerente a diversas temáticas referenciados no conjunto de dados em estudo, ir-se-á primeiramente analisar o mesmo com maior detalhe, e, seguidamente são formados os grupos de gestão responsáveis pelas parcelas.

4.4.1 O conceito de gestão das parcelas

A gestão florestal visa a produção sustentada dos bens e serviços proporcionados pelos recursos existentes na parcela, tendo em conta as atividades e o uso dos espaços envolventes. A administração integrada de parcelas florestais tem como objetivo primordial a definição

e o incremento da produção lenhosa, do aproveitamento dos recursos não lenhosos e outros serviços associados, tendo em consideração a globalidade dos recursos existentes. De facto, a existência de gestão numa parcela está associada à aplicação e administração de uma panóplia de técnicas cujo objetivo primordial consiste no incremento da sua produtividade.

Desta forma, uma parcela foi considerada como sendo gerida, se, à data do inventário, existissem evidências de intervenções de natureza cultural e de exploração dos recursos nas parcelas. Por forma a sustentar as informações recolhidas em campo, estas foram cruzadas com informações existentes na base de dados, referentes aos proprietários da parcela em questão, bem como com fotografias recolhidas em campo.

Por forma a clarificar o conceito de parcela gerida, apresentam-se seguidamente duas imagens de parcelas: com e sem evidências de gestão florestal.

(a)

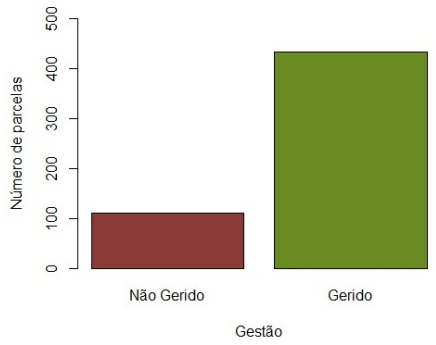
(b)

Figura 4.1: (a) Parcela 26 (área de estudo 2) - não se verifica a existência de gestão. (b) Parcela 138 (área de estudo 1) - verifica-se a existência de gestão.

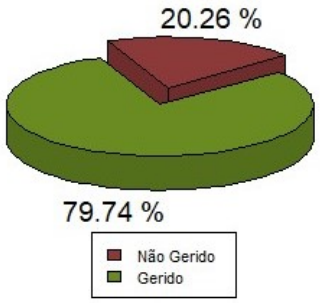
A Figura 4.1a permite verificar que, numa parcela não gerida, a vegetação em torno das árvores não é controlada, não existe qualquer ordenamento na organização das árvores e, apesar da espécie predominante ser o *Eucalyptus Globulus*, esta não é exclusiva, pelo que existem também outras espécies. Antagonicamente, na Figura 4.1b (parcela gerida) verifica-se que a altura da vegetação é mais reduzida, verificando-se o seu controlo; destacando-se ainda um maior cuidado ao nível dos toros das árvores da espécie *Eucalyptus Globulus* na parcela 138.

Após a análise e concatenação das informações relativas à gestão das parcelas, foi possível

concluir a existência de uma proporção de parcelas geridas substancialmente superior à proporção de parcelas não geridas, mormente, 110 parcelas não geridas e 433 parcelas geridas. Tal justifica-se à luz da existência de um maior número de parcelas pertencentes à entidade C - cerca de 377 parcelas- as quais apresentam evidencias de gestão.



(a) Gráfico de barras relativo à contabilização do número de parcelas não geridas e geridas.



(b) Gráfico circular da percentagem de parcelas não geridas e geridas.

Figura 4.2: Contabilização das parcelas geridas e não geridas

4.4.2 Formação dos grupos de gestão no conjunto de dados

O critério estabelecido na formação dos grupos de gestão, consistiu na existência ou inexistência de evidências da aplicação de técnicas de gestão florestal, e na entidade responsável pela parcela, cujas informações foram cruzadas resultando na obtenção dos cinco grupos, a saber:

- $Geridos_A$: formado pelas parcelas geridas (gestão=1) e cuja entidade responsável é a entidade A (Entidade ='A');
- $Geridos_B$: formado pelas parcelas geridas (gestão=1) e cuja entidade responsável é a entidade B (Entidade = 'B');
- $Geridos_C$: formado pelas parcelas provenientes do conjunto de dados **Dados Entidade C**, sendo que neste conjunto de dados todas as parcelas apresentam gestão;
- $NGeridos_A$: formado pelas parcelas não geridas (gestão=0) e cuja entidade responsável e a entidade A (Entidade = 'A');
- $NGeridos_B$: formado pelas parcelas não geridas (gestão=0) e cuja entidade responsável é a entidade B (Entidade = 'B');

Analisando a distribuição das parcelas por tipo de gestão e entidade de forma mais detalhada (consultar figura 4.3 e Anexo C.6), é plausível salientar a inexistência de parcelas classificadas como não geridas pertencentes à entidade C, tal como explanado anteriormente. Com efeito, de entre as parcelas geridas, a entidade C possui um número de parcelas substancialmente

superior ao das restantes entidades, o qual corresponde a cerca de 69.43% do número total de parcelas existentes.

De entre as parcelas em estudo, unicamente parcelas na posse das entidades A e B apresentam evidencias da inexistência de gestão florestal. Relativamente a este grupo de parcelas (não geridas), a maior percentagem pertence à entidade A, o que corresponde a cerca de 91 parcelas, ou seja, 16.76% do número total de parcelas.

Desta forma, as parcelas cuja entidade responsável é a A, é a que possui menor quantidade de parcelas geridas, ou seja, 8 parcelas, o que corresponde a cerca de 1.47% do número total de parcelas. Por outro lado, quanto às parcelas pertencentes à entidade B, constata-se que estas maioritariamente se encontram geridas, sendo que apenas um número reduzido de parcelas, cerca de 28%, não apresenta gestão.

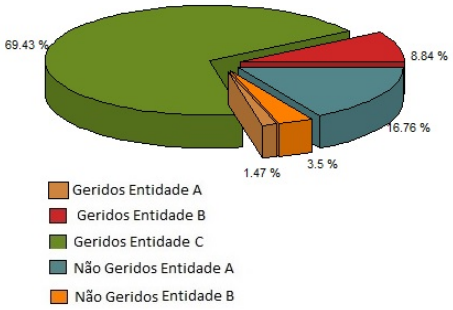


Figura 4.3: Percentagem de parcelas em cada grupo de gestão formado

A análise e contabilização de parcelas em função da gestão aplicada e da entidade responsável permite indagar o número de parcelas em cada grupo de gestão formado, sintetizados na tabela 4.2.

Tabela 4.2: Contabilização do número de parcelas por grupo de gestão formado.

Grupo de Gestão	**Número de parcelas**
$Geridos_A$	8
$Geridos_B$	48
$Geridos_C$	377
$NGeridos_A$	91
$NGeridos_B$	19

4.5 Áreas de estudo

Os inventários florestais foram realizados em duas áreas de estudo, a área de estudo 1 (AE1) e a área de estudo 2 (AE2), pertencentes à zona centro de Portugal Continental, sendo

a primeira mais a norte que a segunda.

É plausível evidenciar a distribuição equitativa dos cinco grupos de gestão pelas duas áreas de estudo, pelo que é verosímil afirmar que o lançamento do inventário florestal desenvolvido pelos investigadores do instituto florestal RAIZ pauta-se por um princípio de equitatividade (Anexo C.7). De facto, na área de estudo 1 contabilizam-se cerca de 300 parcelas em estudo, enquanto que na área de estudo 2 foram contabilizadas cerca de 243 parcelas. Uma análise mais detalhada permite indagar que as parcelas pertencentes ao grupo de parcelas Geridas pela entidade B encontram-se igualmente distribuídas pelas duas áreas de estudo, verificando-se a existência do mesmo número de parcelas deste grupo em cada área de estudo. Também as parcelas geridas pela entidade C se encontram aproximadamente distribuídas de forma equitativa pelas duas áreas de estudo, sendo que a área de estudo 1 possui mais 29 parcelas deste grupo do que a área de estudo 2.

As parcelas cuja entidade responsável é a A, encontram-se distribuídas de forma distinta pelas duas áreas de estudo. Mormente, o número de parcelas geridas pertencentes à referida entidade é substancialmente superior na área de estudo 2, comparativamente com o número destas parcelas existente na área de estudo 1. A referida diferença ronda os 75%. Também no que remete às parcelas não geridas pertencentes à entidade A, constata-se que existem cerca de metade das parcelas na área de estudo 2, comparativamente com o número de parcelas existentes na área de estudo 1.

As diferentes localizações das áreas de estudo, sendo a primeira mais a norte que a segunda, implicam diferentes condições de clima e de solo, e, consequentemente, influenciam a quantidade de matéria prima extraída (isto é, a produtividade da parcela). A Região de Produtividade de uma parcela, RP, conjuga as informações relativas à tipologia de clima e de solo da mesma, sendo que quanto maior a classificação de RP atribuída, menos favorável ao incremento da produtividade é essa região.

Assim, as caixas de bigodes que se encontram no Anexo C.8, sugerem que os valores medianos de AMAutil12, cujos valores fornecem informações relativamente à produtividade das parcelas, são bastante semelhantes entre si nas três regiões de produtividade (RP) em estudo (figura 6 do Anexo C.8).

No âmbito dos valores médios de RP em cada área de estudo, infere-se que em média, a AE 1 é classificada por 3, enquanto que a AE 2 é classificada por 2. A caixa de bigodes da figura 6 (Anexo C.7) revela que a RP3 possui valores de produtividade média superiores aos da RP2, pelo que é plausível concluir que a área de estudo 1 é mais favorável ao incremento da produtividade das parcelas do que a área de estudo 2, no que concerne à respetiva RP, sendo que essas diferenças parecem ser reduzidas.

Por fim, relativamente ao uso de diferentes anos de inventário, verifica-se que tal não é considerado como fator de adulteração dos resultados obtidos nas análises subsequentes na medida em que os valores médios de RP em cada ano de inventário são bastante semelhantes (tabela 4.3). Com efeito, as diferentes condições climáticas verificadas entre 2002 e 2017 não se consideraram fatores adulteradores dos resultados obtidos visto terem assumido valores semelhantes ao longo dos anos.

Tabela 4.3: Valor médio de RP em cada ano inventariado.

Ano	**2002**	**2003**	**2004**	**2005**	**2007**	**2008**	**2009**	**2010**	**2012**	**2013**	**2014**	**2015**	**2017**
RP	2.67	3.2	2.86	3	2.39	2.8	2.4	2.12	2	2	2.66	2.29	2.48

Capítulo 5

Análise dos dados

"A pesquisa científica é um processo de aprendizado dirigido. O objetivo dos métodos estatísticos é tornar este processo o mais eficiente possível"

John Wiley, 1978

A extração de informações preponderantes a partir de um conjunto de dados constitui o cerne da investigação científica ao nível da estatística. Assim, o desenvolvimento de técnicas e metodologias adequadas sob um conjunto de dados permite inferir uma panóplia de conclusões, constituindo um avanço no conhecimento, e induzindo novas perspetivas, estimulando, por sua vez, o ser humano a desenvolver novos conceitos relativos ao mundo em geral.

Ao longo do presente capítulo objetiva-se dar resposta ao problema em estudo. Nesse sentido, foram desenvolvidas inúmeras técnicas, pelo que, primeiramente, foi implementada uma análise exploratória de dados por forma a determinar a(s) metodologia(s) que melhor se coaduna(m). Posteriormente foram aplicadas algumas metodologias de Inferência Estatística, o que, no seguimento da investigação, terá permitido dar resposta à questão colocada pela organização de acolhimento (RAIZ) (formulada na secção 4.1).

A presente investigação fez uso do *software* R (versão 3.4.2) no auxílio da implementação das metodologias estatísticas.

5.1 Análise em Componentes Principais (ACP)

Tendo por base o objetivo da presente investigação, mormente, a análise da influência da gestão florestal na produtividade de plantações de *Eucalyptus Globulus*, e atendendo à dificuldade inerente ao desenvolvimento de um estudo individualizado de cada uma das variáveis indicadoras da produtividade, optou-se por tentar obter uma (ou mais) variáveis que traduzisse(m) a produtividade de uma forma global. Com efeito, a seleção de uma variável relativa à produtividade de uma parcela, poderia ter como consequência a perda de informação pertinente para o presente estudo. Desta forma, recorrendo-se à metodologia de Análise em Componentes Principais (ACP), foi possível agrupar as variáveis detentoras de informação

pertinente no que concerne à produtividade de uma parcela, e assim obter uma ou mais variáveis (objetivando-se idealmente que seja um número inferior às variáveis originais) que fornecem essa informação.

5.1.1 Adequação da ACP

Um dos requisitos da ACP é que as variáveis sobre as quais se objetiva aplicar esta metodologia sejam correlacionadas entre si. As variáveis cuja correlação se considera adequada à aplicação desta metodologia e que detêm informação relativa à produtividade de uma parcela são: área basal das árvores dominantes projetada aos 12 anos (G12), acréscimo médio anual útil projetado aos 12 anos (AMAutil12), volume útil projetado aos 12 anos (Vu12), altura das árvores dominantes projetada aos 12 anos (Hdom12), volume total projetado aos 12 anos (V12) e volume útil mercantil projetado aos 12 anos (Vmudil12). Salienta-se o uso de variáveis cuja informação se encontra projetada aos 12 anos de idade, na medida em que o seu uso permite a comparação de valores entre diferentes parcelas com idades distintas.

A correlação inerente às variáveis supracitadas a partir do gráfico de correlações (figura 5.1) e os gráficos de dispersão entre as variáveis (figura 1 do Anexo D.1) justificam o uso da presente metodologia, visto que se encontram fortemente correlacionadas (considerando valores de correlação superiores a 0.5).

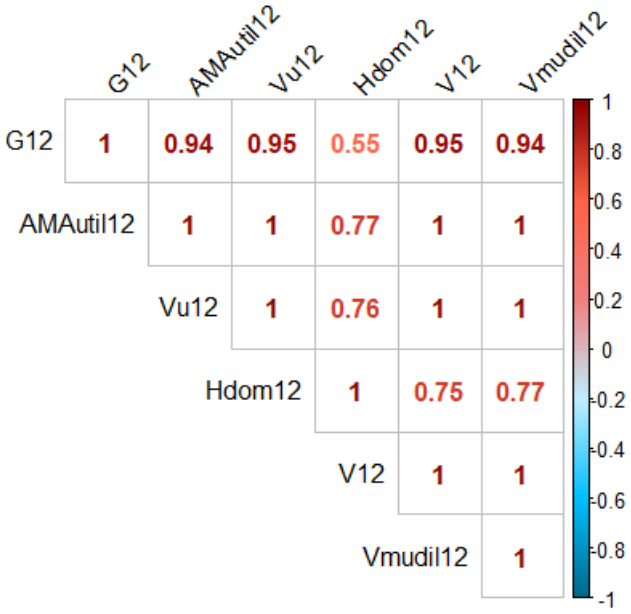


Figura 5.1: Gráfico de correlações (*'corrplot'*) entre as variáveis usadas na ACP.

5.1.2 Número de componentes principais a escolher

A aplicação da ACP, fazendo uso da função *'prcomp()'* disponível no *software* R, resultou na obtenção de tantas componentes principais (CP) quanto o número de variáveis originais usadas; pelo que, tendo a ACP sido aplicada sobre 6 variáveis originais, obtiveram-se 6

componentes principais, formadas a partir de uma combinação linear das variáveis originais, isto é, $CP_i = \sum_{j=1}^{6} \alpha_{ij} v_j$; onde $i = 1, \ldots 6$, v_j designa a j-ésima variável (original), e α_{ij}, designa o valor próprio obtido entre a variável j e a CP_i. Sendo o principal objetivo inerente à aplicação desta metodologia a redução da dimensionalidade, pretende-se, seguidamente adequar o número de componentes principais (CP) a usar, obtendo-se desta forma um menor número destas variáveis, que possibilitam a explicação de uma percentagem apropriada da variabilidade dos dados. Assim, ir-se-á aliar a minimização da perda de informação associada, à redução da dimensionalidade.

Na tabela 5.1 são apresentados os coeficientes das 6 componentes principais obtidas por aplicação de ACP, com recurso ao *software* R. Os valores representados na tabela, designados genericamente por α_{ij}, designam o coeficiente relativo ao peso da variável original j na combinação linear que permite a construção da CP_i.

Tabela 5.1: *Output* do *software* R relativo aos coeficientes das 6 variáveis originais resultantes da aplicação de ACP.

	CP1	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**
$G12$	**0.400**	-0.467	-0.783	0.084	-0.030	1.353e-03
$AMAutil12$	**0.426**	-0.049	0.299	0.479	-0.049	-7.029e-01
$Vu12$	**0.426**	-0.071	0.226	-0.559	-0.670	2.565e-02
$Hdom12$	**0.338**	0.875	-0.348	0.008	-0.002	2.553e-05
$V12$	**0.426**	-0.0837	0.188	-0.478	0.739	-3.411e-02
$Vmudil12$	**0.426**	-0.0496	0.299	0.472	0.011	7.100e-01

Seguidamente efetua-se a decisão do número de componentes a reter com base na escolha de um número de componentes principais que minimize a perda de informação associada ao procedimento. Para tal recorreu-se a três critérios que permitem garantir a consistência do presente procedimento: proporção da variância explicada por cada componente principal, critério de Kaiser e *screeplot*.

a) **Proporção de variância explicada por cada componente principal:**
O presente critério recorre à quantidade de variância explicada pelas componentes principais. Assim, segundo este critério, o número de componentes principais a escolher é o que explica, pelo menos, cerca de 85% da variabilidade total, [34].

Tabela 5.2: Proporção de variância explicada por cada componente principal.

	CP1	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**
Proporção de variância	**0.915**	0.081	0.004	0.0002	0.000	0.000
Proporção de variância acumulada	**0.915**	0.996	0.999	1.000	1.000	1.000

A análise da tabela 5.2 permite constatar que, considerando a escolha de um número de componentes principais que permite explicar pelo menos 85% da variabilidade total

das variáveis originais, a primeira componente principal explica aproximadamente 91% da variabilidade total das variáveis originais, pelo que a sua escolha é adequada.

O gráfico subsequente coaduna-se com o explanado anteriormente, permitindo verificar visualmente a adequação da escolha da primeira componente principal segundo o critério em estudo.

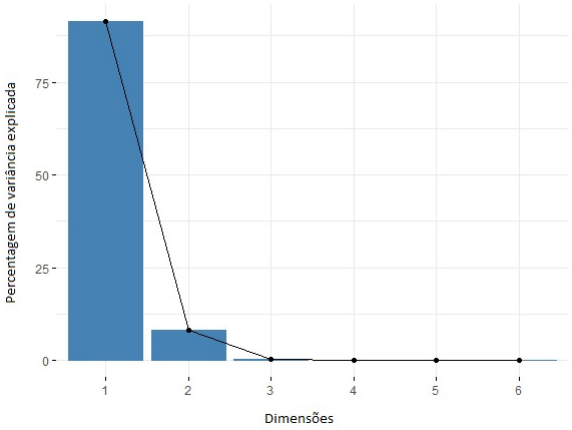


Figura 5.2: Gráfico da percentagem de variância explicada por cada CP.

b) **Critério de Kaiser**
Este critério, proposto por Kaiser em 1960, sugere que se devem reter as componentes principais cujos valores próprios sejam superiores a um, [34]. No presente estudo constata-se que, atendendo à tabela 5.3, apenas a primeira componente principal possui valor próprio superior a 1, e, por conseguinte, o presente critério sustenta a escolha da primeira componente principal.

Tabela 5.3: Valores próprios de cada CP, para aplicar o critério de Kaiser.

	CP1	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**
Valores Próprios	**2.343**	0.697	0.155	0.031	0.005	0.003

O gráfico da figura 5.3 permite confirmar a escolha da primeira componente principal, com base no critério de Kaiser, uma vez que a reta horizontal interceta a curva representada a azul no valor próprio da primeira CP.

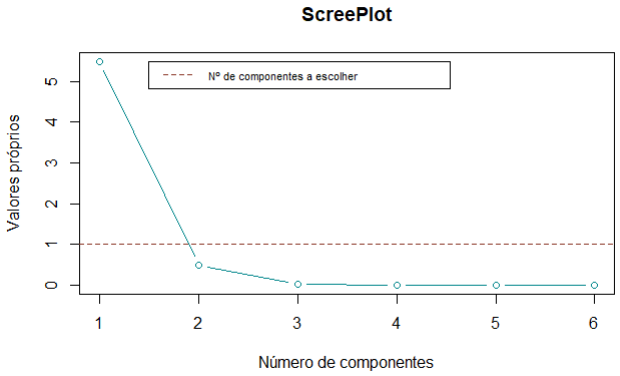


Figura 5.3: Gráfico que relaciona os valores próprios com a respetiva CP - aplicação visual do critério de Kaiser

c) **Screeplot - "gráfico do cotovelo"**
Este critério, proposto por Cattell em 1966, [48], sugere que o número de componentes principais a reter deve ser aquele que antecede o momento em que o declive atenua, ou seja, quando os valores próprios das componentes principais forem próximos entre si e de zero. O *screeplot* obtido na aplicação da ACP ao conjunto de dados em estudo (figura 5.4) corrobora com as conclusões indagadas pela aplicação dos critérios supracitados. Assim, a análise visual do *screeplot* permite intuir a seleção de todas as componentes até ao ponto de inflexão da curva, isto é, até que a linha comece a ficar aproximadamente horizontal, o que leva à escolha apenas da primeira componente principal.

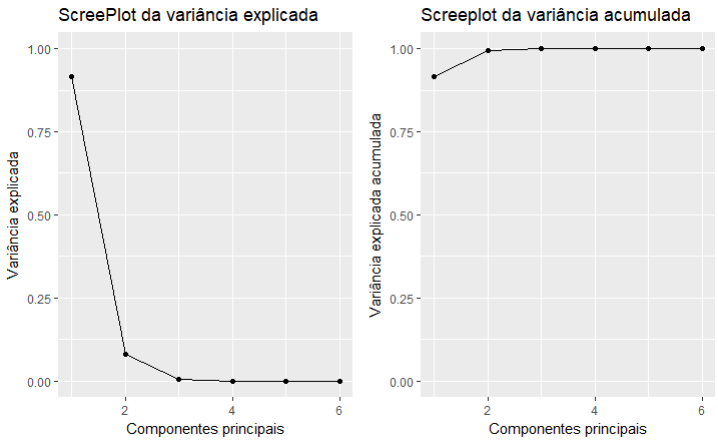


Figura 5.4: Dois *screeplots*, sendo que à esquerda se relaciona a quantidade de variância, e à direita a quantidade de variância acumulada explicadas para as CPs obtidas.

Tendo por base a aplicação simultânea dos três critérios, ambos permitem depreender a escolha da primeira componente principal, permitindo a referida escolha explicar cerca de 91% da variabilidade total das variáveis originais.

À primeira componente principal selecionada, designou-se **produtividade** na medida em que resulta da combinação linear de variáveis que colecionam informação relativa à produ-

tividade da parcela. Assim, os valores dos coeficientes da primeira componente principal relativamente às variáveis originais encontram-se sintetizados na tabela seguinte.

Tabela 5.4: Coeficientes da primeira componente principal relativamente às variáveis originais - *output* obtido pelo *software* R.

Variáveis Originais	**CP1**
$G12$	0.402
$AMAutil12$	0.426
$Vu12$	0.426
$Hdom12$	0.337
$V12$	0.426
$Vmudil12$	0.426

Com efeito, a variável produtividade resulta da seguinte combinação linear:

$$\begin{aligned} produtividade = {} & 0.402G12 + 0.426Amautil12 + 0.426Vu12 \\ & + 0.337Hdom12 + 0.426V12 + 0.426Vmudil12 \end{aligned} \tag{5.1}$$

5.1.3 Características sumárias da variável produtividade

Com o intuito de analisar com maior detalhe as características sumárias da variável produtividade, obtida por aplicação de ACP, sintetizam-se os resultados na tabela 5.5. As características sumárias da variável produtividade, obtidas recorrendo à função *'summary()'* do *software* R, permitem verificar que a produtividade mínima é de 17.78 unidades de medida, o que corresponde à parcela 192, a qual pertence ao grupo de parcelas não geridas pertencentes à entidade A. Por outro lado, é plausível intuir que a produtividade máxima é na ordem de 761.65 unidades, correspondendo à parcela 2428, na qual se verifica a existência de evidências de gestão florestal aplicada pela entidade C. A produtividade média é na ordem de 237.42 unidades.

Tabela 5.5: Características sumárias da variável produtividade.

Mínimo	**1ºQuartil**	**Mediana**	**Média**	**3ºQuartil**	**Máximo**
17.78	159.13	225.05	237.42	301.37	761.65

5.2 Adequação dos testes paramétricos

O principal objetivo inerente à presente investigação consiste na comparação da produtividade em parcelas com e sem gestão, à qual se alia a informação relativa às três entidades responsáveis pelas mesmas – entidades A, B e C. Tendo por base os fundamentos teóricos apresentados na secção 3.2, a utilização de testes potentes conduz à escolha preferencial de testes paramétricos e, consequentemente, à necessidade de proceder à validação dos seus

pressupostos. Com efeito, esse procedimento requer que se verifiquem as seguintes condições, de forma concomitante: normalidade da produtividade da parcela; e homogeneidade das suas variâncias.

A análise da adequação da aplicação de testes paramétricos aos dados é feita, numa primeira fase, sob os grupos de parcelas geridas e não geridas, cuja análise se encontra no Anexo D.2. As análises permitiram concluir que, devido à falta de Normalidade inerente aos conjuntos de parcelas geridas e não geridas, é ilícita a aplicação de testes paramétricos aos grupos supracitados.

Seguidamente são analisadas as condições de aplicabilidade dos testes paramétricos aos grupos formados na secção 4.4.2.

5.2.1 Análise da normalidade

A análise de normalidade dos grupos formados na secção 4.4.2, baseia-se, numa primeira fase, na observação gráfica conjunta da função densidade de probabilidade estimada da variável produtividade para cada grupo de gestão, e posteriormente, para a confirmação dos resultados, procedeu-se à aplicação de testes de Normalidade. Assim com recurso ao *software* R, foram obtidos os gráficos das estimativas das funções densidades de probabilidade referentes à variável produtividade nos cinco grupos de gestão formados (figura 5.5 e figuras 4 e 5 do Anexo D.3). A sua construção permite refletir os padrões gerais da função densidade de probabilidade que se objetiva estudar na presente subsecção, sendo que, a sua análise se prende com a necessidade de colocar em evidencia a existência ou inexistência de tendências de Normalidade inerentes aos grupos em estudo.

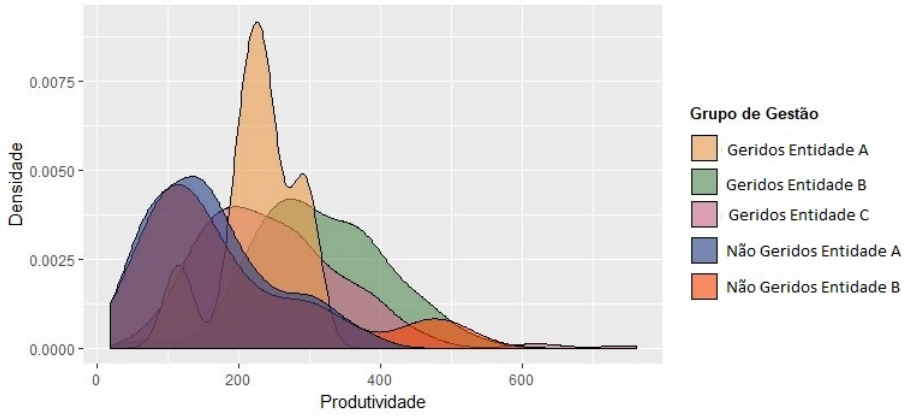


Figura 5.5: Esboço da função densidade estimada da variável produtividade para os cinco grupos de gestão formados.

A curva da função densidade de probabilidade da distribuição Normal caracteriza-se por ser simétrica, possuindo a forma de sino. A partir das representações gráficas da função densidade de probabilidade estimada da variável produtividade nos diferentes grupos, é plausível indagar que apenas as curvas referentes à função densidade de probabilidade estimada dos grupos de parcelas geridas pelas entidades A e B evidenciam simetria (aproximada), visto que ambas possuem uma forma aproximadamente semelhante a um sino.

Com o auxílio do *software* R, fez-se uso das funções '*qqplot()*' e '*hist()*', as quais permitiram obter, respetivamente, o *qqplot* e os histogramas, para a variável produtividade, para cada um dos grupos formados na secção 4.4.2, por forma a inferir visualmente, de forma mais detalhada, quanto à Normalidade da variável produtividade em cada grupo. Seguidamente implementaram-se os testes de normalidade aplicados aos grupos de dados estabelecidos na secção 4.4.2, por forma a comprovar os resultados obtidos pelas representações gráficas elaboradas anteriormente. Nesse sentido recorreram-se aos testes: **Shapiro-Wilk**, **Anderson-Darling** e teste de **Kolmogorov-Smirnov com correção de Lilliefords**. Note-se que neste caso o teste de Kolmogorov-Smirnov (sem correção de Lilliefords) considerou-se desadequado tendo em conta que os valores da média e do desvio padrão associados à amostra em estudo eram desconhecidos. A escolha dos testes de Normalidade aplicados baseou-se nas características inerentes ao conjunto de dados em estudo e na adequação dos respetivos testes à dimensionalidade e às características dos mesmos. A aplicação conjunta dos testes mencionados permite comprovar os resultados obtidos individualmente. Recorrendo ao *software* R, faz-se uso das funções '*shapiro.test()*', '*ad.test()*' e '*lillie.test()*', respetivamente.

Salienta-se, por fim, a necessidade de submeter à avaliação da Normalidade todas as categorias e subamostras sujeitas à análise e não apenas a amostra global, atendendo a que em estudos subsequentes se pretende analisar e efetuar comparações múltiplas entre os subgrupos de parcelas geridas e não geridas e entre os subgrupos formados.

Primeiramente, far-se-á uma análise ao grupo de parcelas geridas pertencentes à entidade A, cujo grupo é formado por um número reduzido de parcelas: oito.

Os dois gráficos obtidos com recurso ao *software* R - *qqplot* e o histograma - referentes à variável produtividade para o grupo de parcelas geridas pertencentes à entidade A, permitem intuir a existência de Normalidade nos dados. Com efeito, os valores dos quantis calculados a partir da amostra dos dados relativos à produtividade das parcelas geridas pela entidade A aproximam-se dos quantis teóricos da distribuição Normal para a presente amostra, na medida em que os pontos (valores empíricos) se aproximam da reta do *qqplot* (valores teóricos da distribuição Normal). Também em relação ao histograma, verifica-se que existe um pico de frequências, e uma distribuição aproximadamente simétrica das restantes barras em torno desse pico, sendo plausível indagar a aproximação do histograma à curva a vermelho relativa aos valores teóricos da distribuição Normal.

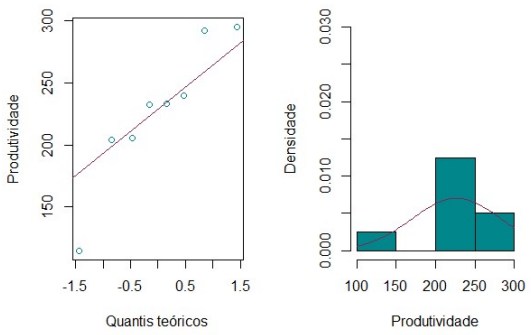


Figura 5.6: À esquerda encontra-se representado o *qqplot*, e à direita o histograma e respetiva curva da distribuição Normal, relativos à função densidade da variável produtividade no grupo de parcelas geridas pertencentes à entidade A.

A conjetura apresentada requer confirmação recorrendo a testes de hipóteses de Normalidade, não só devido à falta de precisão e rigor inerentes às representações gráficas, mas também devido à reduzida dimensionalidade do grupo em estudo, o qual é formado apenas por 8 parcelas.

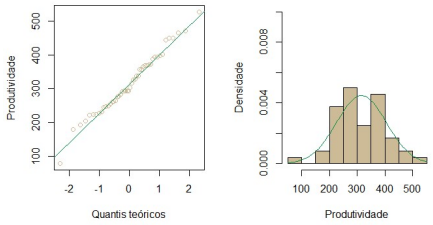
A hipótese nula dos testes aplicados pressupõe a existência de Normalidade da variável produtividade relativamente ao grupo de parcelas geridas pertencentes à entidade A. Assim, a um nível de significância de 5%, todos os testes aplicados corroboram a não rejeição da hipótese nula, o que está de acordo com as conclusões indagadas graficamente. Conclui-se, assim, a existência de Normalidade da variável produtividade relativamente ao grupo das parcelas geridas pertencentes à entidade A.

Tabela 5.6: Resultados dos testes de Normalidade para o grupo de parcelas geridas pertencentes à entidade A, relativamente à sua produtividade.

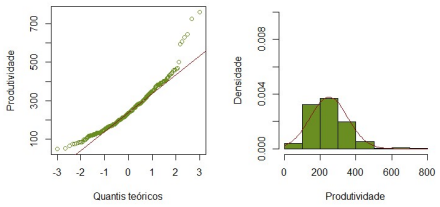
	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Shapiro-Wilk	0.338	Não rejeitar normalidade
Teste de Anderson-Darling	0.307	Não rejeitar normalidade
Teste de KS com correção de lilliefors	0.538	Não rejeitar normalidade

Analogamente, os resultados obtidos para os restantes grupos de gestão formados são apresentados seguidamente.

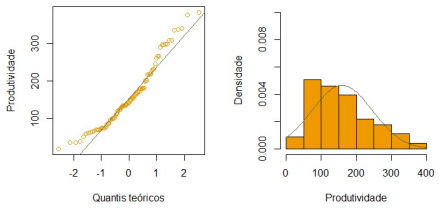
No que infere às representações gráficas apresentadas, de forma análoga ao que foi explanado anteriormente, permitem conjeturar a aproximação à distribuição Normal no que infere ao grupo de parcelas geridas cuja entidade responsável é a entidade B. Por outro lado, o afastamento evidenciado entre os *qqplots* e os histogramas empíricos das respetivas representações teóricos, permitem supor a falta de Normalidade dos restantes grupos de parcelas. As representações gráficas encontram-se na figura 5.7.



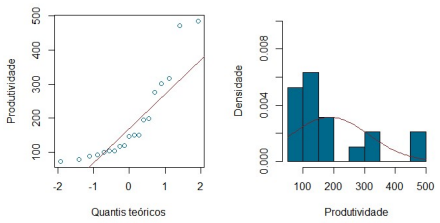
(a) Geridas entidade B.



(b) Grupo entidade C.



(c) Não Geridas entidade A.



(d) Não geridas entidade B

Figura 5.7: *Qqplot*, e histograma relativos à função densidade estimada da variável produtividade nos respetivos grupos de gestão formados.

Por forma a comprovar os resultados conjeturados graficamente, são implementados testes de Normalidade, nomeadamente os testes de **Shapiro-Wilk**, **Anderson-Darling** e **KS com correção de Lilliefors**, cujos resultados se encontram sintetizados na tabela 5.7.

Tabela 5.7: Resultados dos testes de Normalidade para os grupos de gestão formados, relativamente à variável produtividade.

		p-value	**Decisão** ($\alpha = 0.05$)
Geridas Entidade B	**Teste de Shapiro-Wilk**	0.921	Não rejeitar normalidade
	Teste de Anderson-Darling	0.765	Não rejeitar normalidade
	Teste de KS com correção de lilliefors	0.430	Não rejeitar normalidade
Geridas C	**Teste de Shapiro-Wilk**	0.0004	Rejeitar normalidade
	Teste de Anderson-Darling	0.0003	Rejeitar normalidade
	Teste de KS com correção de lilliefors	0.012	Rejeitar normalidade
Não Geridas Entidade A	**Teste de Shapiro-Wilk**	0.008	Rejeitar normalidade
	Teste de Anderson-Darling	0.006	Rejeitar normalidade
	Teste de KS com correção de lilliefors	0.047	Rejeitar normalidade
Não Geridas Entidade B	**Teste de Shapiro-Wilk**	0.002	Rejeitar normalidade
	Teste de Anderson-Darling	0.001	Rejeitar normalidade
	Teste de KS com correção de lilliefors	0.009	Rejeitar normalidade

Note-se que o grupo de parcelas geridas pertencentes à entidade C possui 377 parcelas, pelo que neste caso o teste de Shapiro-Wilk não se considera adequado tendo em conta a elevada dimensionalidade inerente ao grupo supracitado. Por outro lado, sendo o grupo de parcelas não geridas pertencentes à entidade B formado por 19 parcelas, o teste que melhor se coaduna com a dimensionalidade inerente aos dados é o teste de Shapiro-Wilk, apesar dos resultados dos restantes testes aplicados permitirem obter resultados semelhantes.

Os resultados dos testes de Normalidade permitem assim concluir, a um nível de significância de 5% a não rejeição da hipótese de Normalidade no grupo de parcelas geridas pertencentes à entidade B. Não obstante, relativamente ao mesmo nível de significância, os testes implementados permitem rejeitar a hipótese de Normalidade nos grupos de parcelas geridas pela entidade C, e nos grupos de parcelas não geridas pertencentes às entidades A e B.

5.2.2 Análise da homogeneidade das variâncias

Uma outra condição que é necessário verificar para se coadunar a aplicação de um teste de hipóteses paramétrico é a de que as variâncias populacionais dos grupos a comparar sejam homogéneas.

Os testes aplicados foram o **teste de Levene** e o **teste de Fligner**. Ambos os testes são classificados como dois dos testes mais potentes utilizados para este fim, sendo considerados testes particularmente robustos a desvios de Normalidade, [52], pelo que a sua aplicação se adequa ao conjunto de dados em estudo.

Assim, a análise da homogeneidade dos grupos formados na secção 4.4.2, inicia-se pela sua aplicação sob os grupos de parcelas geridas pelas três entidades. Consequentemente, a hipótese nula de ambos os testes pressupõe a existência de homogeneidade da variância da variável produtividade entre os três grupos de parcelas geridas cujas entidades responsáveis podem ser A, B ou C.

Com efeito, sendo σ_i^2 a variância da variável produtividade, relativa ao grupo i ($i = Geridos_A, Geridos_B, Geridos_C$) a hipótese nula traduz-se por:

$$H_0 : \sigma^2_{Geridos_A} = \sigma^2_{Geridos_B} = \sigma^2_{Geridos_C}$$

Com o auxílio do *software* R, foi possível implementar os testes de Levene e de Fligner, tendo-se recorrido para tal às funções '*levene.test()*' e '*fligner.test()*', respetivamente. Os resultados obtidos encontram-se sintetizados na tabela 5.8.

Tabela 5.8: Resultados obtidos no que concerne à homogeneidade das variâncias nos grupos de parcelas geridas pelas 3 entidades (A, B e C) - testes de Levene e de Fligner.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Levene	0.098	Não rejeitar homogeneidade
Teste de Fligner	0.076	Não rejeitar homogeneidade

A hipótese nula dos testes de Levene e de Fligner assumem, ambos, a existência de homogeneidade das variâncias dos três grupos de parcelas geridas. Assim, tendo em conta os resultados sintetizados na tabela 5.8 é verosímil afirmar que, sendo o *p-value* superior a 0.05, não existem evidencias para rejeitar a hipótese nula a um nível de significância de 5%. Por conseguinte, não existem motivos para rejeitar a igualdade das variâncias dos três grupos das parcelas geridas pertencentes às três entidades - A, B e C. É de realçar, no entanto, que se fossem considerados outros níveis de significância habitualmente utilizados, por exemplo, $\alpha = 0.1$, a decisão seria a de rejeição da hipótese nula, isto é, a rejeição da homogeneidade das variâncias nos três grupos, relativamente à produtividade.

Analogamente, implementou-se o mesmo procedimento considerando as parcelas não geridas, pertencentes às entidades A e B. Neste caso a hipótese nula sob teste traduz-se por:

$$H_0 : \sigma^2_{NGeridos_A} = \sigma^2_{NGeridos_B},$$

onde σ_i^2 representa a variância da produtividade, relativamente ao grupo i, onde i designa os grupos em questão. Os resultados obtidos sintetizam-se na tabela 5.9.

Tabela 5.9: Resultados obtidos no que infere à homogeneidade das variâncias nos grupos de parcelas não geridas pelas 2 entidades (A e B) - testes de Levene e de Fligner.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Levene	0.2664	Não rejeitar homogeneidade
Teste de Fligner	0.5425	Não rejeitar homogeneidade

A implementação dos testes supracitados permite concluir a não rejeição da homogeneidade das variâncias da produtividade ao nível dos grupos de parcelas não geridas, pertencentes às entidades A e B, tendo por base a obtenção de valores de *p-value* superiores a 0.05, que implicam a não rejeição da hipótese nula, a um nível de significância de 0.05.

5.2.3 Conclusão

Tendo em conta a violação da normalidade nos grupos formados, mormente ao nível do grupo de parcelas geridas cuja entidade responsável é C, apesar de se verificar homogeneidade das variâncias da variável produtividade de entre os grupos de parcelas geridas, não é possível a aplicação de testes paramétricos, pelo que surge a necessidade de usar testes não paramétricos na presente investigação, cuja adequação se analisa mais detalhadamente na secção 5.3.

5.3 Adequação dos testes não paramétricos

No presente capítulo justifica-se de forma mais detalhada, o uso dos testes não paramétricos na comparação de medianas populacionais da produtividade entre os grupos de gestão formados, implementados nas secções subsequentes. A sua utilização justifica-se com base nas seguintes razões:

a) **Falta de Normalidade nos grupos formados**, nomeadamente, no grupo de parcelas geridas pela entidade C, parcelas não geridas pertencentes a entidade B e parcelas não geridas pertencentes à entidade A (analisado na secção 5.2.1);

b) Alguns grupos de gestão formados possuem um número de parcelas reduzido, em particular: os grupos de parcelas geridas pertencentes à entidade A (formado por 8 parcelas) e de parcelas não geridas pertencentes à entidade B (constituído por 19 parcelas).

Os testes não paramétricos aplicados na sequência da presente investigação foram: **teste de Kruskal-Wallis**, **teste de Wilcoxon-Mann-Whitney** e **ANOVA a dois fatores não paramétrica**.

5.4 Análise conjunta da influência da gestão e da entidade na produtividade das parcelas

A presente secção prende-se com a necessidade de efetuar uma análise comparativa da produtividade entre parcelas que apresentam evidências de gestão florestal e parcelas que não apresentam evidências de gestão florestal, ou seja, parcelas pertencentes às entidades A e B. Para tal recorreu-se a ANOVA a dois fatores. Tendo em conta as características inerentes ao conjunto de dados em estudo, nomeadamente a inexistência de parcelas não geridas pertencentes à entidade C – o que impossibilita a comparação com base no fator gestão - a presente metodologia foi aplicada apenas aos dados pertencentes às entidades A e B.

Assim, a aplicação do método ANOVA a dois fatores (*two-way*) não paramétrica objetiva investigar a influência concomitante de dois fatores na produtividade da parcela - **gestão** e **entidade**. Desta forma, a presente metodologia pretende testar o efeito do fator gestão da parcela e também a interação entre a gestão e a entidade responsável pela parcela, na produtividade da mesma. A inexistência de Normalidade na variável produtividade nos grupos de parcelas não geridas pertencentes às entidades A e B, instigou a aplicação da presente metodologia não paramétrica, em detrimento da sua versão paramétrica.

A aplicação da presente metodologia permite dar resposta às questões:

1. A gestão (ou a entidade) tem ou não um efeito significativo sobre a produtividade da parcela;

2. Se o efeito da entidade responsável pela parcela sobre a produtividade da mesma é ou não influenciado pela existência ou inexistência de gestão.

Primeiramente, com base num gráfico de interação entre os fatores, foi possível detetar a existência de uma relação entre os fatores Gestão e Entidade na produtividade mediana das parcelas, visível pela inexistência de paralelismo entre as retas do gráfico da figura 5.8. As caixas de bigodes da referida figura permitem também evidenciar a associação existente entre as parcelas geridas e valores de produtividade mais elevados, e as parcelas não geridas, e valores de produtividade mais reduzidos (figura 5.8).

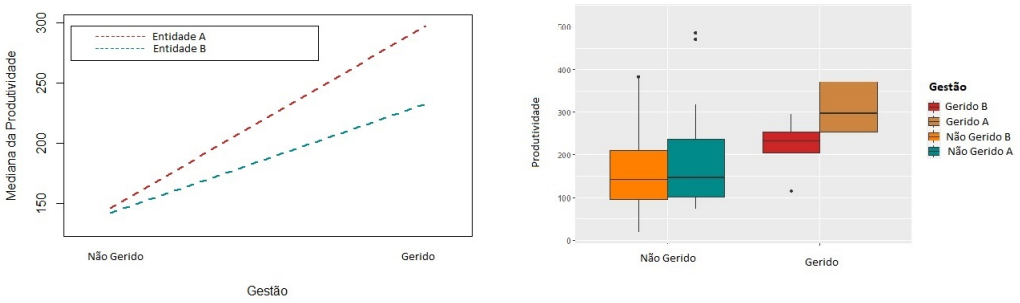


Figura 5.8: À esquerda encontra-se o gráfico da interação entre os fatores, e à direita as caixas de bigodes dos grupos de parcelas geridas e não geridas pelas entidades A e B.

Ir-se-á, seguidamente, confirmar a conjetura supracitada recorrendo à ANOVA a dois fatores não paramétrica. As hipóteses sob teste podem ser escritas como:

H_0^{Gestao}: O fator Gestão não tem um efeito significativo na produtividade da parcela.

vs

H_1^{Gestao}: O fator Gestão tem um efeito significativo na produtividade da parcela.

$H_0^{Entidade}$: O fator Entidade não tem um efeito significativo na produtividade da parcela.

vs

$H_1^{Entidade}$: O fator Entidade tem um efeito significativo na produtividade da parcela.

H_0^{γ}: Não existe interação entre os fatores Gestão e Entidade responsável pela parcela.

vs

H_1^{γ}: Existe interação entre os fatores Gestão e Entidade responsável pela parcela.

Para testar cada uma das hipóteses supracitadas é necessário calcular a correspondente estatística de teste H para o fator Gestão, para o fator Entidade, e para a interação $\gamma = Gestao \times Entidade$.

Assim, primeiramente começou por se efetuar a ordenação de todas as observações relativas à variável produtividade por ordem crescente, mantendo a identificação da origem da observação relativamente aos fatores em estudo, mormente relativamente à Gestão e à Entidade, sendo que aos empates foram atribuídas as ordens médias. A ordenação foi implementada no *software* R com recurso à função *'rank()'*.

Seguidamente, tendo por base os conceitos teóricos inerentes à presente metodologia (secção 3.3.3), é necessário efetuar os cálculos relativos aos valores de $SQOF$ para ambos os fatores, Gestão e Entidade, e relativamente à sua interação, e o valor de $QMOT$, por forma a calcular o valor da estatística de teste H.

As fórmulas relativas ao cálculo da estatística de teste H e respetivos *p-values* foram implementadas numa função do *software* R, permitindo assim generalizar a sua utilização para outros conjuntos de dados (consultar Anexo D.4). Uma alternativa à função desenvolvida, seria recorrer à função *'anova()'* do *software* R, aplicada aos dois fatores, gestão e entidade, e sob os dados da produtividade ordenados (usando, de forma análoga a função *'rank()'*). Os resultados obtidos por ambos os métodos supracitados são análogos.

A função implementada necessita primeiramente da especificação dos parâmetros. Seguidamente far-se-á uma breve exposição do seu significado, e respetivos valores na amostra em estudo.

- **Ordem:** vetor previamente ordenado, neste caso relativo à variável produtividade, sobre a qual se pretende avaliar a influência dos dois fatores. A ordenação da variável recorre

ao comando '*rank()*' do *software* R;

- **FA:** Vetor relativo à variável que designa o primeiro fator, neste caso, é o vetor relativo à gestão das parcelas;
- **FB:** Vetor relativo à variável que designa o segundo fator, neste caso, é o vetor que contém informação relativa às entidades responsáveis pela parcela;
- **a:** número de níveis do primeiro fator, neste caso, é o número de níveis do fator gestão, ou seja, é 2 (Sim ou Não);
- **b:** número de níveis do segundo fator, neste caso, é o número de níveis do fator entidade, ou seja, é 2 (entidade A ou entidade B);
- **n:** número de conjugações entre os níveis de ambos os fatores em estudo. Neste caso é 4 pois é possível fazer as seguintes combinações Entidade/Gestão: A/Sim; A/Não; B/Sim; B/Não;
- **N:** número de observações existentes no conjunto de dados.
- **Alfa:** nível de significância que se pretende usar no teste, sendo que neste caso foi considerado um nível de significância de 5%;

A especificação dos referidos parâmetros na função implementada, permite a obtenção dos resultados sintetizados na tabela 5.10.

Com base na metodologia apresentada na secção 3.3.3, conclui-se que se rejeita H_0 com uma probabilidade de erro do tipo I não superior a α se $H \geqslant \chi^2_{1-\alpha,(g.l.)}$. Assim, adicionalmente, na função construída foram implementadas condições que, a cada fator e respetiva interação, permitem indagar quanto à rejeição da hipótese nula. Tal permite facilitar a leitura e interpretação dos resultados obtidos.

Tabela 5.10: Resultados obtidos por aplicação de ANOVA a dois fatores não paramétrica aos dados sem a entidade C, comparando os fatores gestão e entidade, relativamente à variável produtividade.

Origem da variação	**Soma dos quadrados das ordens**	**Graus de liberdade**	**H**	***p-value***	**Decisão** ($\alpha = 0.05$)
Fator Gestão	10850995	1	4697.061	$< 2e-16$	Rejeitar H_0
Fator Entidade	10912441	1	4723.659	0.006	Rejeitar H_0
Interação entre Gestão e Entidade	-5009884	1	-2168.624	0.650	Não rejeitar H_0
Total	16753552	3			

Assim, para um nível de significância de 5%, é plausível concluir a rejeição das hipóteses nulas que assumem que a gestão (isoladamente) e a entidade (isoladamente) não influenciam a produtividade de uma parcela. Assim, quer a gestão quer a entidade consideradas isoladamente

influenciam a produtividade das parcelas. Não obstante, não se rejeita a hipótese de que a interação é nula. Ou seja, não existem evidências que permitam rejeitar que a inexistência de interação entre o tipo de gestão florestal aplicado e a entidade responsável pela parcela influenciam a produtividade da mesma.

A título conclusivo, constata-se que se encontram reunidas as condições necessários para responder às questões colocadas no inicio da presente subsecção:

1. **A gestão (ou a entidade) tem ou não um efeito significativo sobre a produtividade da parcela?**
   Tendo em conta o valor reduzido do *p-value* obtido (próximo de zero) para o efeito da gestão na produtividade da parcela, tal implica a rejeição da hipótese nula. Com efeito, rejeita-se que a gestão aplicada na parcela, por si só, não influencia a produtividade da mesma. Assim, a gestão, isoladamente, influencia a produtividade das parcelas. Adicionalmente, conclui-se que as conclusões indagadas para a entidade responsável pela parcela são análogas: a entidade responsável pela parcela, por si só, influencia a produtividade da mesma, visto que o valor do *p-value* reduzido (aproximadamente nulo) levou à rejeição da hipótese nula de que a entidade não influencia a produtividade das parcelas.

2. **Se o efeito da entidade responsável pela parcela sob a produtividade da mesma é ou não influenciado pela existência ou inexistência de gestão?**
   A interação entre a gestão florestal desenvolvida e a entidade responsável pela parcela pode não produzir um efeito significativo na produtividade da mesma, visto que, o valor do *p-value* superior a 0.05, implica, a um nível de significância de 5%, a não rejeição de que a interação entre os fatores gestão e entidade é nula.

Tendo em conta o explanado anteriormente (secção 3.3.3), verifica-se que a aplicação da presente metodologia se mostrou profícua, permitindo explicar a realidade de forma fidedigna, tendo por base a obtenção de *p-values* que induzem a rejeição da hipótese nula de inexistência de influência entre cada um dos fatores isoladamente na produtividade das parcelas, e, em contrapartida, não é induzida a rejeição da referida hipótese para a influência da interação dos fatores na produtividade das parcelas.

Assim, tendo por base as conclusões supracitadas, mormente a não rejeição de que a interação entre os fatores gestão e entidade é nula, é plausível estudar os referidos fatores separadamente, e respetiva influência na produtividade das parcelas.

5.5 Influência da gestão na produtividade

Seguidamente avalia-se de que forma a produtividade de uma parcela é influenciada pela existência ou inexistência de gestão na mesma.

Com o intuito de desenvolver uma sensibilização primária, construiu-se uma caixa de bigodes entre os dois grupos de parcelas (geridas e não geridas) e a produtividade das mesmas, apresentada na figura 5.9.

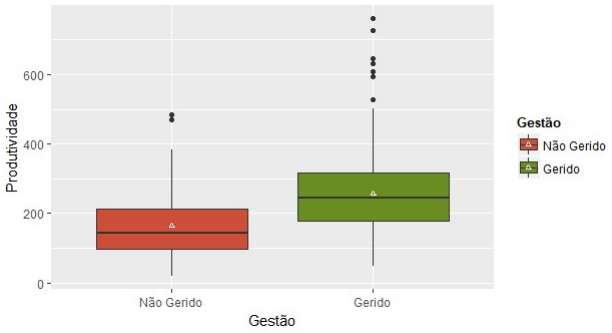


Figura 5.9: Caixa de bigodes entre a existência/inexistência de gestão e a produtividade das parcelas.

A figura 5.9 sugere que a produtividade das parcelas geridas é incrementada com a existência de gestão, na medida em que a caixa de bigodes relativa às parcelas geridas encontra-se acima da caixa de bigodes relativa às parcelas não geridas. Com efeito, analogamente à caixa de bigodes, a figura 5.10 sugere que a produtividade média nas parcelas geridas é superior à produtividade média em parcelas não geridas, visível pela localização do pico da estimativa da função densidade da produtividade para valores mais elevados desta variável. De facto, a produtividade média de parcelas geridas é, aproximadamente, 256.000 unidades de medida, enquanto que a produtividade média de parcelas não geridas, ronda 164.282 unidades de medida. A figura 5.9 sugere ainda que os valores da produtividade mínima nos dois grupos são próximos entre si, sendo que a produtividade mínima é mais reduzida em parcelas não geridas. De facto, a produtividade mínima em parcelas geridas é 48.81 unidades de medida, enquanto que em parcelas não geridas assume o valor de 17.78 unidades de medida. Por outro lado, a produtividade máxima é superior em parcelas geridas (aproximadamente 761.65 unidades de medida), do que em parcelas não geridas (aproximadamente, 485.48 unidades).

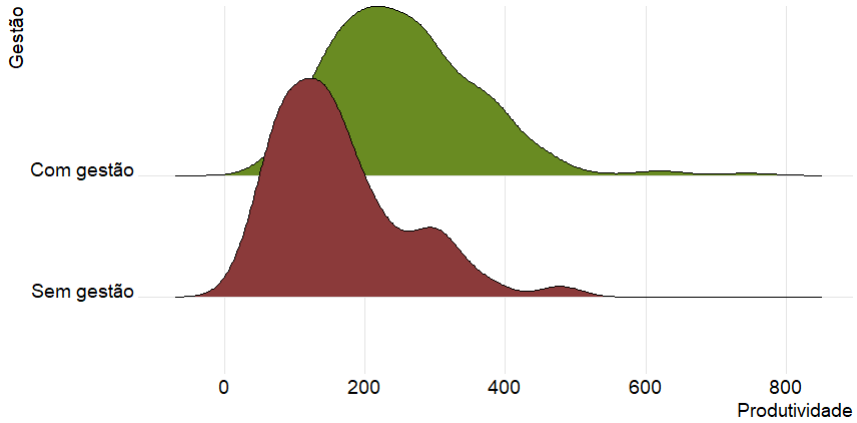


Figura 5.10: Gráfico da função densidade estimada da variável produtividade por existência/inexistência de gestão nas parcelas.

Nas parcelas onde não existem evidências da existência de gestão florestal constata-se que o valor mediano da produtividade é inferior ao das parcelas que apresentam evidencias da aplicação de técnicas de gestão florestal.

Por forma a comprovar as conclusões decorrentes da análise gráfica, foram aplicados testes não paramétricos para comparar a produtividade entre os dois grupos de parcelas: geridas e não geridas. A inadequação dos testes paramétricos aos dados, tendo por base a falta de normalidade inerente aos grupos de parcelas em estudo ao nível da variável produtividade (Anexo D.2), motivou o uso do teste não paramétrico de Wilcoxon-Mann-Whitney. Os resultados obtidos pela aplicação unilateral do referido teste comprovam as conjeturas mencionadas relativas à análise gráfica. Com efeito, a sua aplicação permite concluir que o valor mediano da produtividade do grupo de parcelas Geridas é significativamente superior ao respetivo valor no grupo de parcelas Não Geridas (visto que o *p-value* do teste de Wilcoxon-Mann-Whitney unilateral à direita implementado foi inferior a 2.2×10^{-16}). Assim, a maioria das parcelas geridas possui valores de produtividade superiores aos das parcelas que não apresentam gestão florestal. Desta forma, o valor mediano em parcelas não geridas e em parcelas geridas, é, respetivamente, aproximadamente, 143.657 e 244.244 unidades de medida.

No que concerne à variabilidade da produtividade nos grupos de parcelas Geridas e Não Geridas, a caixa de bigodes da figura 5.9 permite intuir que a variabilidade interquartis da produtividade nas parcelas geridas é ligeiramente superior à variabilidade interquartis da produtividade nas parcelas que não apresentam evidências da existência de gestão.

Da aplicação dos testes de Levene e Fligner para a homogeneidade de variâncias, obtiveram-se os p-values (ver tabela 2 do Anexo D.2) 0.3131 e 0.257, respetivamente. Consequentemente, não se rejeita que a variação da produtividade seja igual nas parcelas geridas e nas parcelas não geridas, a um nivel de significância de 0.05.

5.6 Influência dos grupos de gestão na produtividade das parcelas

Seguidamente, pretende-se estudar se os grupos de gestão (formados na secção 4.4.2) influenciam a produtividade das parcelas.

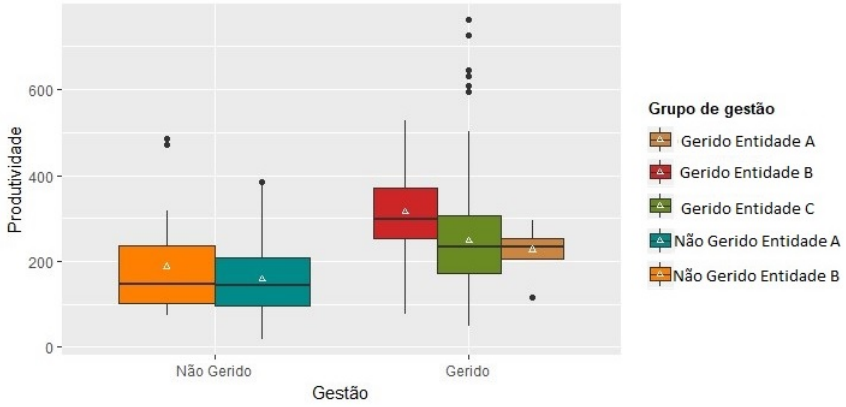


Figura 5.11: Caixa de bigodes entre os grupos de gestão e a respetiva produtividade das parcelas.

Uma análise preliminar sob a caixa de bigodes da figura 5.11 sugere que o grupo cuja

produtividade é superior é o grupo de parcelas que apresentam evidências de gestão, pertencentes à entidade B.

No que concerne à variabilidade interquartis entre a produtividade das parcelas não geridas, a caixa de bigodes apresentada na figura 5.11 sugere a existência de variabilidades interquartis semelhantes entre as parcelas pertencentes à entidade A, e pertencentes à entidade B. Tal é visível a partir do afastamento semelhante existente entre os valores do 1º e do 3º quartis das caixas de bigodes referentes às parcelas não geridas pertencentes às entidades A e B. Tal constatação corrobora com as conclusões indagadas por aplicação dos testes de Levene e de Fligner (na secção 5.2.2, tabela 5.9), onde a obtenção de um valor do *p-value* superior a 0.05, a um nível de significância de 5%, implica a não rejeição da hipótese nula, e, por conseguinte, é plausível concluir a não rejeição da homogeneidade das variâncias da produtividade entre os dois grupos de parcelas não geridas, pertencentes às entidades A e B.

No que diz respeito à variabilidade da produtividade entre os três grupos de parcelas geridas, a análise das caixas de bigodes da figura 5.11 instiga que, sendo as caixas de bigodes referentes às parcelas pertencentes à entidade B e à entidade C do mesmo tamanho, a variabilidade da produtividade nestes dois grupos de parcelas é igual. Salienta-se o tamanho reduzido da caixa de bigodes referente às parcelas geridas cuja entidade responsável é a entidade A o que se justifica com base no reduzido número de parcelas pertencentes a esse grupo (8 parcelas). Os testes da homogeneidade das variâncias apresentados na secção 5.2.2, corroboram com as conjeturas supracitadas, visto que sustentam a igualdade das variâncias da produtividade entre os três grupos de parcelas geridas.

Salienta-se que, relativamente ao valor médio da produtividade da parcela, cujo valor se encontra representado por um ponto branco nas caixas de bigodes apresentadas, as conjeturas são semelhantes às indagadas no âmbito da mediana da produtividade. Com efeito, conjetura-se assim que, o valor médio da produtividade é superior em parcelas geridas pela entidade B, sendo inferior em parcelas não geridas pertencentes à entidade A.

Grupos de parcelas sem gestão

A análise das caixas de bigodes apresentadas na figura 5.11 permite intuir que, relativamente às parcelas que não apresentam evidencias de gestão, os valores medianos da produtividade nos grupos de parcelas pertencentes às entidades A e B são semelhantes.

Por forma a comprovar as conjeturas formuladas, e tendo por base a inadequação do uso de testes paramétricos na presente comparação (analisada na secção 5.2), recorreu-se ao teste de Wilcoxon-Mann-Whitney na comparação das medianas da produtividade entre os dois grupos em estudo. A obtenção de um *p-value* superior a 0.05 (tabela 5.11), permite intuir que, relativamente às parcelas onde não existem evidencias da aplicação de metodologias de gestão florestal cujas entidades responsáveis são a entidade A ou a entidade B, as medianas relativas à variável produtividade para os dois grupos não são significativamente diferentes, a um nível de significância de 5%.

Tabela 5.11: Resultados obtidos por aplicação do teste de Wilcoxon-Mann-Whitney aos grupos de parcelas não geridas - pertencentes às entidades A e B, onde θ representa a mediana da produtividade das parcelas.

Hipóteses do teste de Wilcoxon-Mann-Whitney	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0: \ \theta_{NaoGeridos_A} = \theta_{NaoGeridos_B}$ ***vs*** $H_1: \ \theta_{NaoGeridos_A} \neq \theta_{NaoGeridos_B}$	0.3551	Não rejeitar igualdade das medianas

Porém, verifica-se que a produtividade assume valores mínimos mais reduzidos em parcelas não geridas pertencentes à entidade A, do que em parcelas não geridas pertencentes à entidade B. Tal é visível a partir da localização inferior do valor mínimo da caixa de bigodes referente às parcelas não geridas pertencentes à entidade A, em detrimento do valor relativo à caixa de bigodes relativa às parcelas não geridas pertencentes à entidade B. De facto, o valor mínimo da produtividade nas parcelas não geridas pertencentes à entidade A é 17.78 unidades, enquanto que nas parcelas não geridas pertencentes à entidade B é 72.655.O valor máximo da produtividade é inferior em parcelas não geridas pertencentes à entidade A (aproximadamente 383.578) do que o respetivo valor referente às parcelas não geridas pertencentes à entidade B (aproximadamente 485.485).

Grupo de parcelas com gestão

É plausível intuir que, inerente à gestão aplicada pelas 3 entidades, são aplicadas metodologias de gestão florestal distintas, pelo que seguidamente objetiva-se investigar de que forma os diferentes tipos de gestão aplicados pelas três entidades influenciam a produtividade das parcelas. Assim, no que concerne às parcelas onde predominam evidencias da existência de gestão florestal, a caixa de bigodes apresentada na figura 5.11 instiga que, o valor da mediana da produtividade é superior em parcelas pertencentes à entidade B, seguida pelas parcelas geridas pela entidade C, e por fim, assumindo valores inferiores em parcelas pertencentes à entidade A, cujo valor é bastante próximo ao correspondente valor da entidade C.

As conclusões supracitadas são comprovadas recorrendo a testes não paramétricos, tendo em conta a inadequação da aplicação de testes paramétricos aos grupos de parcelas em estudo pelas razões apresentadas na secção 5.2.

Primeiramente aplicou-se o teste de Kruskal-Wallis, com o intuito de comparar o comportamento da variável produtividade nos três grupos de parcelas geridas pelas três entidades A ($Geridos_A$), B ($Geridos_B$) e C ($Geridos_C$), tornando assim possível testar se existem diferenças significativas na mediana da produtividade resultante por aplicação de gestão florestal pelas três entidades supracitadas. Nesse sentido, recorre-se à comparação da mediana da variável produtividade nos três grupos de gestão, pelo que as hipóteses implementadas no teste de Kruskal-Wallis traduzem-se por:

$$H_0 : \theta_{Geridos_A} = \theta_{Geridos_B} = \theta_{Geridos_C}$$

vs

$$H_1 : \exists i, j : \ \theta_i \neq \theta_j \quad (i \neq j = Geridos_A, Geridos_B, Geridos_C)$$

onde θ designa a mediana da produtividade da população. Os resultados encontram-se sintetizados na tabela 5.12.

Tabela 5.12: Resultados obtidos por aplicação do teste de Kruskal-Wallis entre os grupos de parcelas onde há gestão.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Kruskal-Wallis	2.555×10^{-5}	Rejeitar igualdade das medianas nos três grupos

Tendo por base a hipótese nula do teste de Kruskal-Wallis, e os resultados obtidos através da sua aplicação (sintetizados na tabela 5.12) , sendo os valores do *p-value* muito reduzidos, tomando valores inferiores a 0.05, existem evidências que permitem rejeitar a hipótese nula a um nível de significância de 5%. Consequentemente conclui-se que as parcelas geridas pelas três entidades, A, B e C não provêm da mesma população, relativamente à variável produtividade.

A conclusão de que existe pelo menos um grupo de gestão formado cujo valor da mediana da variável produtividade é diferente dos restantes grupos, seguidamente ir-se-á averiguar qual o grupo que contribui para essa diferença. Nesse sentido, aplica-se seguidamente o teste de Wilcoxon-Mann-Whitney (unilateral) aos grupos de parcelas geridas entre as três entidades. Os resultados obtidos encontram-se sintetizados na tabela 5.13.

Tabela 5.13: Resultados obtidos por aplicação do teste de Wilcoxon-Mann-Whitney aos grupos de parcelas com evidências de gestão pertencentes às três entidades de gestão - A, B e C, onde θ representa a mediana da produtividade das parcelas.

Hipóteses do teste de Wilcoxon-Mann-Whitney	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0 : \ \theta_{Geridos_A} = \theta_{Geridos_C}$ ***vs*** $H_1 : \ \theta_{Geridos_A} \neq \theta_{Geridos_C}$	0.728	Não rejeitar igualdade das medianas
$H_0 : \ \theta_{Geridos_C} = \theta_{Geridos_B}$ ***vs*** $H_1 : \ \theta_{Geridos_C} < \theta_{Geridos_B}$	3.007×10^{-6}	Rejeitar igualdade das medianas
$H_0 : \ \theta_{Geridos_A} = \theta_{Geridos_B}$ ***vs*** $H_1 : \ \theta_{Geridos_A} < \theta_{Geridos_B}$	0.002581	Rejeitar igualdade das medianas

A hipótese nula do teste de Wilcoxon-Mann-Whitney assume que as medianas em comparação são iguais. Assim, a um nível de significância de 5%, os resultados obtidos preconizam que a gestão efetuada pela entidade C permite a obtenção de valores medianos da produtividade inferiores aos da gestão aplicada pela entidade B. A gestão efetuada por esta entidade permite ainda a obtenção de valores medianos da produtividade superiores aos da gestão implementada

pela entidade A. Não obstante, a mediana da produtividade em parcelas cuja gestão aplicada é da responsabilidade da entidade C pode ser considerada similar à mediana da produtividade em parcelas cuja gestão é efetuada pela entidade A.

As conjeturas formuladas sob os valores médios da produtividade entre os grupos de gestão das parcelas a partir da análise gráfica das caixas de bigodes são análogas.

Assim, a análise dos gráficos das figuras 5.11 e 5.12, e ainda as conclusões indagadas a partir dos testes implementados, permite intuir a existência de uma produtividade decrescente no sentido: $Geridos_B$, $Geridos_C$, $Geridos_A$, $NaoGeridos_B$ e $NaoGeridos_A$.

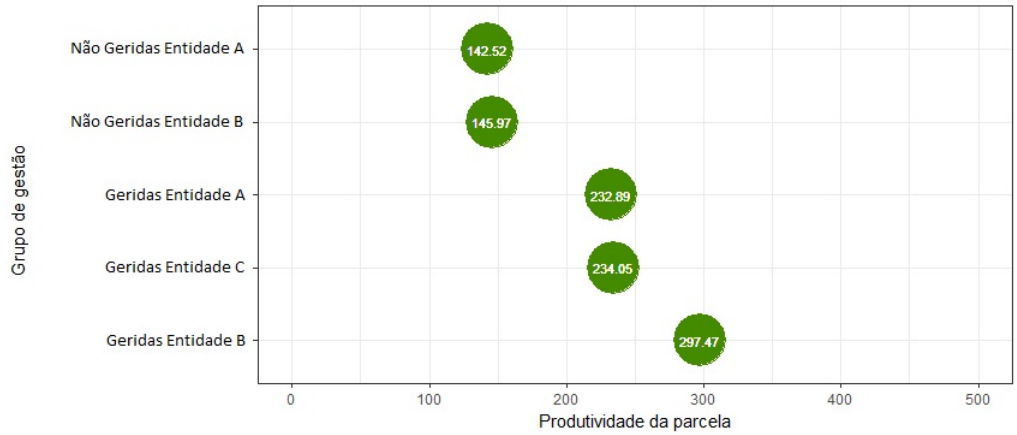


Figura 5.12: Gráfico divergente de valores medianos - mediana da variável produtividade por grupo de gestão.

Assim, a título conclusivo é plausível intuir que, de uma forma geral, a produtividade é superior em parcelas Geridas pela entidade B e apresenta valores mais reduzidos em parcelas cuja entidade responsável é a entidade A.

5.7 Influência de outros fatores na gestão e produtividade das parcelas

Nas secções anteriores foram implementadas análises no sentido de avaliar a influência da gestão e da entidade responsável pela parcela na sua produtividade. Não obstante, verifica-se que, para além da gestão das parcelas, existe uma panóplia de variáveis que influenciam a sua produtividade. Assim, na presente secção objetiva-se tomar consciência da relação entre algumas variáveis consideradas de relevo, mormente, a idade das parcelas, e algumas metodologias de gestão florestal, na produtividade das mesmas.

5.7.1 A idade da parcela

A informação relativa à idade da parcela é primordial na estimação e previsão da quantidade de matéria prima aí existente, resultando do cruzamento de informação obtida por inquirição local, com informações da base de dados.

As parcelas com plantações da espécie *Eucalyptus Globulus* possuem maior quantidade de matéria prima usada para fins de produção de papel entre os 8 e os 12 anos, idade a partir da qual as árvores da referida espécie são cortadas, garantindo de forma concomitante, um crescimento tanto maior quanto possível e a otimização da quantidade de matéria prima obtida, cuja informação é comprovada tendo por base a literatura consultada, [51, 28], e as análises estatísticas implementadas. Assim, nas parcelas em estudo a idade máxima verificada é 14 anos.

Seguidamente analisa-se de que forma a gestão e os grupos de gestão influenciam a idade das parcelas, e, por sua vez, de que forma a idade das parcelas influencia a produtividade das mesmas.

5.7.1.1 Relação entre gestão e idade das parcelas

Primeiramente ir-se-á analisar a relação entre a idade e a existência ou inexistência de gestão florestal numa parcela, com base no gráfico de bigodes da figura 5.13.

A sua análise revela que uma parcela onde sejam aplicadas técnicas de gestão florestal apresenta evidencias, no que concerne à idade da mesma, que sugerem a aplicação de técnicas e cuidados florestais que inibem a existência de parcelas com elevada amplitude de idades das árvores. Antagonicamente, em parcelas onde não se verifica a aplicação de técnicas de gestão florestal, predomina um leque de idades maior, na medida em que os cuidados são substancialmente reduzidos. Tal é visível a partir da diferença acrescida entre os quartis da idade das parcelas não geridas relativamente ao seu valor no grupo de parcelas geridas. Os testes relativos à homogeneidade da variância das idades nos grupos de parcelas geridas e não geridas, passiveis de serem consultados na tabela 3 do Anexo D.5, mormente os valores reduzidos do *p-value* (inferiores a 0.05), os quais permitem concluir a inexistência de homogeneidade da variância entre os grupos em estudo, o que se coaduna com as conclusões deduzidas pela observação das caixas de bigodes. Consequentemente é plausível concluir a existência de um leque de idades superior em parcelas não geridas, relativamente ao leque de idade das parcelas geridas.

A análise das caixas de bigodes sugere ainda que o valor mediano da idade das parcelas geridas é ligeiramente superior ao da idade das parcelas não geridas. As conclusões supracitadas comprovam-se recorrendo ao teste não paramétrico de Wilcoxon-Mann-Whitney (unilateral à esquerda) cuja implementação permitiu obter um *p-value* de 0.0022, aproximadamente, pelo que, a um nível de significância de 5%, conclui-se a rejeição da hipótese nula de igualdade das variâncias. Consequentemente, intui-se a superioridade do valor da mediana da idade das parcelas pertencentes ao grupo de parcelas geridas, em detrimento do seu valor no grupo de parcelas não geridas.

Os dois grupos de parcelas partilham os mesmos valores máximo e mínimo de idades das parcelas.

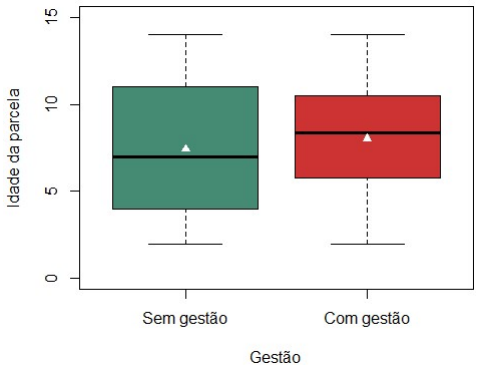


Figura 5.13: Caixas de bigodes entre parcelas com e sem gestão e respetiva idade.

A entidade responsável pela parcela, e, por conseguinte, o grupo de gestão a que pertencem (formados na secção 4.4.2), deverão também influenciar de forma significativa a sua idade. Assim, seguidamente ir-se-á analisar essa relação, com base na figura 5.14.

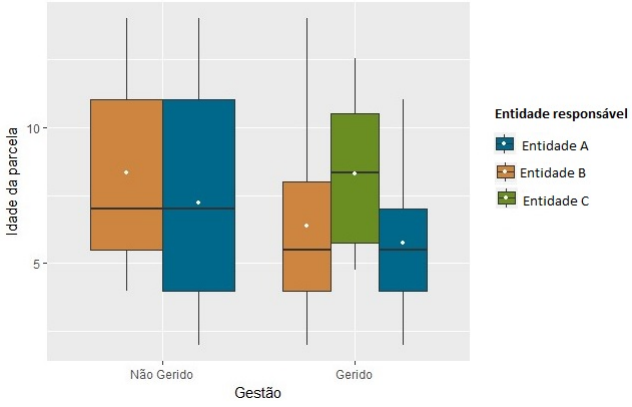


Figura 5.14: Caixas de bigodes entre parcelas dos grupos formados e respetiva idade.

Relativamente aos grupos de parcelas que não apresentam evidencias de gestão, cujas entidades responsáveis são as entidades A ou B, é plausível constatar que o comportamento da idade das parcelas no que infere à sua variação é, aproximadamente, semelhante. A conclusão conjeturada por observação gráfica corrobora com as conclusões inferidas a partir dos testes de

hipóteses da homogeneidade das variâncias entre os grupos supracitados (consultar resultados na tabela D.5 do Anexo D.5). A obtenção de valores de *p-value* superiores a 0.05 permite depreender que não há evidências para rejeitar a hipótese nula, que assume a homogeneidade das variâncias nos dois grupos em estudo.

Relativamente ao grupo de parcelas geridas, as caixas de bigodes apresentadas sugerem que a amplitude de idades das parcelas é distinta entre as três entidades de gestão florestal. As conclusões supracitadas são comprovadas recorrendo aos testes de homogeneidade das variâncias, cujos resultados se encontram na tabela D.5 do Anexo D.5. Estes permitem comprovar a existência de homogeneidade das variâncias entre os grupos de parcelas $Geridos_A$ e $Geridos_B$, e entre os grupos de parcelas $Geridos_A$ e $Geridos_C$. Não obstante, deteta-se a inexistência de homogeneidade das variâncias entre as parcelas pertencentes aos grupos $Geridos_B$ e $Geridos_C$.

As caixas de bigodes apresentadas na figura 5.14 permitem ainda presumir que o valor mediano da idade das parcelas geridas pela entidade C é substancialmente superior ao valor mediano da idade das parcelas geridas pelas restantes entidades, sendo que as parcelas pertencentes às entidades A e B possuem a mesma mediana da idade. As conclusões supracitadas comprovam-se recorrendo ao teste de Wilcoxon-Mann-Whitney, cujos resultados se encontram na tabela D.5 do Anexo D.5.

Relação entre idade e produtividade das parcelas

A idade de uma parcela destaca-se no incremento da quantidade de matéria prima extraída da parcela, e, por conseguinte, na produtividade da mesma. De facto, a idade de corte das árvores de uma parcela influencia substancialmente a quantidade de matéria prima extraída (e consequentemente a sua produtividade), visto que a espécie *Eucalyptus Globulus* possui maior quantidade de matéria prima útil em determinadas idades. Com o intuito de avaliar essa relação, foi construído o gráfico da figura 5.15.

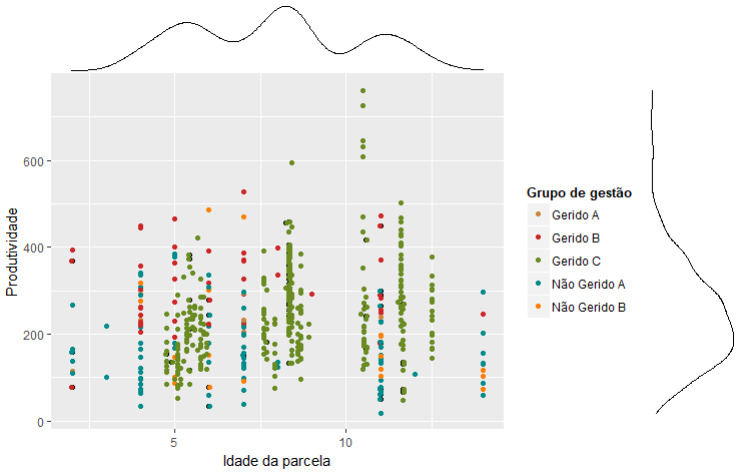


Figura 5.15: Gráfico de dispersão e respetiva função densidade das variáveis idade e produtividade.

De uma forma geral, a figura 5.15 preconiza a existência de uma maior panóplia de valores

de idades em parcelas não geridas pertencentes à entidade A, visível pela maior dispersão dos pontos referentes a este grupo. O referido gráfico sugere que a produtividade nestas parcelas é mais reduzida, comparativamente com a produtividade dos restantes grupos.

O gráfico de dispersão apresentado na figura 5.15 permite deduzir ainda que a produtividade de uma parcela é maximizada entre os 8 anos e os 12 anos, mormente, em parcelas geridas pela entidade C.

É ainda possível conjeturar que, à medida que a idade da parcela aumenta, a produtividade aumenta também, verificando-se que os valores mais elevados de produtividade se verificam em parcelas geridas pela entidade B. Por outro lado, as parcelas cuja produtividade é mais reduzida são não geridas pertencentes à entidade A.

5.7.2 As metodologias de gestão florestal

A seleção de varas das árvores existentes na parcela, o controlo da vegetação da parcela, e o tipo de preparação do terreno constituem exemplos de metodologias de gestão florestal comummente usadas nas parcelas em estudo. A aplicação das técnicas supracitadas constitui uma evidencia que favorece a existência de gestão na parcela. Na presente subsecção pretende-se relacionar as metodologias supracitadas com o grupo de gestão e a produtividade da parcela.

Seleção de varas

A seleção de varas consiste no corte das árvores a uma determinada altura. Após o corte as toiças voltam a rebentar, sendo que a partir de cada toiça desenvolvem-se várias varas que devem ser selecionadas dois a três anos após o primeiro corte, para que, dessa seleção, resultem as varas mais vigorosas.

No âmbito da aplicação da presente metodologia pelos grupos de gestão formados, e tendo por base a figura 6 do Anexo D.6, é plausível deduzir que a aplicação de técnicas de seleção de varas é implementada de forma mais regular em parcelas onde existem evidencias da existência de gestão, mormente, em parcelas geridas pela entidade C, cuja totalidade das parcelas se pautam pela aplicação desta técnica. Verifica-se que nas parcelas não geridas pertencentes à entidade A, maioritariamente não se verifica seleção de varas. Salienta-se ainda que, contrariamente ao que seria expectável por se tratarem de parcelas geridas, este grupo de parcelas (pertencentes à entidade A) não apresentam seleção de varas.

A análise da caixa de bigodes da figura 5.16 permite conjeturar que a mediana da produtividade é superior em parcelas geridas onde não se verifica a aplicação desta metodologia.

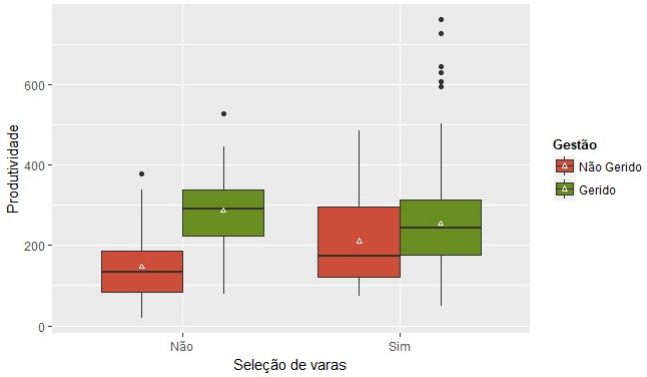


Figura 5.16: Caixas de bigodes entre a aplicação de seleção de varas e a produtividade, relativamente ao grupo de gestão das parcelas.

A conclusão supracitada relativa à produtividade das parcelas comprova-se recorrendo ao teste não paramétrico de Wilcoxon-Mann-Whitney. Os resultados obtidos por aplicação do referido teste podem ser consultados na tabela 6 do Anexo D.6, permitindo comprovar que a mediana da produtividade é superior em parcelas geridas onde não se verifica seleção de varas, seguida do grupo de parcelas geridas onde existe seleção de varas, sendo que a diferença entre os dois grupos supracitados é reduzida. As parcelas não geridas onde não existe aplicação de seleção de varas são as que possuem valores de mediana da produtividade mais reduzidos. Note-se que, tendo por base a literatura consultada,[51], seria expectável que as parcelas geridas onde a presente metodologia fosse aplicada permitiria obter resultados mais aliciantes ao nível da produtividade das parcelas. Não obstante, tal não ocorre, sendo que uma justificação plausível para tal baseia-se nas idades reduzidas das parcelas geridas, cuja aplicação da presente metodologia não se coaduna. Consequentemente, a decisão da aplicação desta metodologia deve ter em consideração a idade da parcela.

Preparação do terreno

A Preparação do terreno consiste numa metodologia fundamental para o crescimento das árvores, na medida em que visa melhorar o arejamento, infiltração, retenção de água e disponibilidade de nutrientes no solo, as quais constituem condições essenciais ao desenvolvimento radicular, crescimento das plantas e sucesso das plantações de *Eucalyptus Globulus*. Com efeito, a promoção da manutenção e/ou aumento da matéria orgânica do solo, e o consequente ajuste das técnicas às características do solo, permitem associar a preparação do terreno à garantia da qualidade ambiental e económica dos projetos, gerindo de forma adequada a biomassa verde ou seca existentes no terreno.

A aplicação das diferentes técnicas de preparação do terreno depende não só do tipo de solo, mas também do grupo de gestão responsável pela parcela. Os gráficos que sustentam a análise supracitada encontram-se na figura 7 do Anexo D.7. Assim, quanto às parcelas geridas, verifica-se que, nas que pertencem à entidade C, o tipo de preparação do terreno implementada apenas varia entre três técnicas: "sem armação" (a qual é aplicada em exclusivo nas parcelas pertencentes a este grupo), "terraços" e "vala e cômoro". De forma análoga, a preparação do terreno aplicada nas parcelas geridas pela entidade A varia entre quatro

técnicas – cava, desalinhado, ripagem e terraços. Nas parcelas cuja entidade responsável é a entidade B, verifica-se que as metodologias de preparação do terreno são bastante diversificadas, independentemente de existirem ou não evidencias de gestão florestal.

As diferentes técnicas de preparação do terreno aplicadas objetivam não só adequar a plantação de *Eucalyptus Globulus* à tipologia de solo, mas também favorecer a produtividade da parcela, pelo que a seguir se irá estudar essa relação. A observação da caixa de bigodes apresentada na figura 5.17 sugere que, de uma forma geral, o valor mediano da produtividade é superior nas parcelas geridas, para cada tipo de preparação do terreno aplicado, cujos resultados são sustentados com base nos resultados do teste de Wilcoxon-Mann-Whitney (passível de ser consultado na tabela 8 do Anexo D.7). Ao nível das parcelas onde existem evidencias de gestão florestal, as caixas de bigodes sugerem que a mediana da produtividade é superior quando a preparação do terreno aplicada é 'linha', sendo mínima caso a preparação do terreno aplicada seja 'sem armação'.

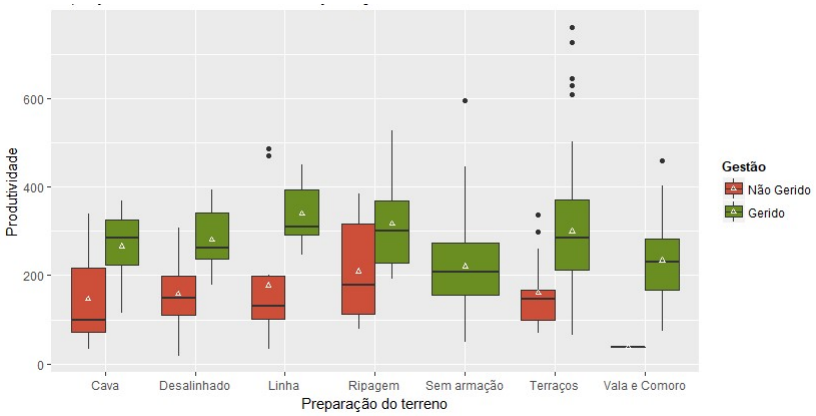


Figura 5.17: Caixas de bigodes entre o tipo de preparação do terreno aplicado nas parcelas, e a sua produtividade, em função da existência ou inexistência de gestão nas parcelas.

Controlo da vegetação

O controlo da vegetação numa parcela é primordial no incremento da produtividade da mesma, visto que a vegetação espontânea compete com os eucaliptos por água, luz e nutrientes, o que pode levar à redução da produtividade das plantações.

O controlo da vegetação é aplicado de forma preferencial em parcelas onde existem evidências de gestão florestal. Nesse sentido, de entre os grupos de parcelas em estudo, apenas as que pertencem à entidade A não possuem controlo da vegetação (independentemente da existência ou inexistência de gestão florestal). A totalidade das parcelas pertencentes às entidades B e C caracterizam-se pela existência de controlo da vegetação das parcelas (independentemente da existência ou inexistência de gestão florestal). As conclusões supracitadas são sustentadas a partir da figura 8 do Anexo D.8.

As caixas de bigodes apresentadas na figura 5.18 sugerem que o controlo da vegetação permite incrementar a mediana da produtividade das parcelas que apresentam evidências de gestão, o que é visível tendo em conta a localização superior do valor da mediana da produtividade do grupo de parcelas geridas com controlo da vegetação, relativamente ao das

restantes caixas de bigodes. A conclusão supracitada comprova-se recorrendo ao teste não paramétrico de Wilcoxon-Mann-Whiney, cuja aplicação permite obter os resultados sintetizados na tabela 10 do Anexo D.8. Com efeito, foi possível constatar que, nas parcelas onde não existiam evidências da aplicação de gestão florestal, a mediana da produtividade foi superior em parcelas com controlo da vegetação, em detrimento de parcelas onde não se verificava o referido controlo.

Assim, aliar o controlo da vegetação e outras metodologias de gestão florestal apresenta-se como primordial no incremento da produtividade das parcelas.

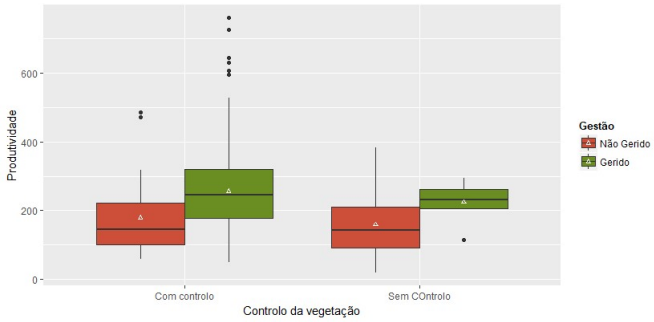


Figura 5.18: Caixas de bigodes entre a existência ou inexistência de controlo de vegetação e a produtividade das parcelas, segundo as evidências de gestão aí existentes

5.8 Regressão

Com o intuito de verificar e indagar a existência de outros fatores que influenciam a produtividade das parcelas (para além da gestão e da entidade vistas anteriormente), bem como o sentido em que essa influência ocorre, e respetiva dimensão, efetua-se seguidamente uma análise de regressão. O principal objetivo inerente é verificar de que forma, as diferentes metodologias aplicadas pelas três entidades responsáveis pelas parcelas influenciam a sua produtividade. Com efeito, a aplicação da presente metodologia possibilita a avaliação da influência da aplicação de diferentes metodologias de gestão florestal pelas três entidades, conjugando essa informação com outros fatores de relevo, na produtividade das parcelas.

O presente estudo baseia-se assim no conjunto de dados composto apenas pelas parcelas geridas, pertencentes às três entidades supracitadas. Na sequência do explanado, a variável dependente/resposta considerada na aplicação da presente metodologia é a produtividade da parcela. De entre as variáveis pertencentes ao conjunto de dados, importa considerar, numa primeira fase, as seguintes variáveis independentes/explicativas, cujos motivos da sua escolha se prendem com análises exploratórias efetuadas e com a respetiva definição, as quais permitem intuir quanto à influência dessas variáveis na produtividade das parcelas:

- frequência do número de árvores vivas (FR_{nvivas});
- idade validada da parcela ($Idade_{validada}$);
- rotação (Rotação);
- seleção de varas (SelecVaras);
- controlo da vegetação ($Controlo_{Vegetacao}$);
- região de produtividade da parcela (RP);
- entidade responsável pela parcela ($entd$).

Primeiramente, implementou-se regressão linear (utilizando variáveis quantitativas e qualitativas), na qual se considerou como variável dependente a produtividade, e as variáveis independentes supracitadas, tendo-se feito uso da função '*lm()*' do *software* R, que, por omissão, estima os parametros de regressão pelo método dos mínimos quadrados.

Os resultados obtidos encontram-se sintetizadas na tabela 5.14, a partir dos quais se conclui que nem todas as variáveis são significativas no modelo, visto que nem todos os coeficientes associados às variáveis independentes do modelo de regressão são significativamente diferentes de zero, e por conseguinte, nem todos possuem valores do *p-value* significativos (inferiores a 0.05), considerando um nível de significância de 5%.

Tabela 5.14: Resultados obtidos por implementação do modelo de regressão linear '*lm()*'sem seleção de variáveis pelo método *Stepwise*, a partir do *software* R.

	Estimativa do coeficiente	**Erro Padrão**	***t-value***	***p-value***
Intercept	169.056	29.997	5.636	3.19e−08***
Fr_{nvivas}	114.957	31.558	3.643	0.0003***
Idade validada	4.710	1.994	2.362	0.019*
Rotação (talhadia)	-61.516	12.623	-4.873	1.56e−06***
AE(2)	-17.653	10.702	-1.649	0.099
Seleção de varas(Sim)	-23.750	36.932	-0.643	0.521
Controlo da vegetação(sem controlo)	225.519	112.816	1.999	0.046 *
RP(3)	65.643	19.451	3.375	0.000807***
RP(4)	25.200	12.511	2.014	0.045*
Entidade (B)	51.955	26.169	1.985	0.048*
Entidade (A)	-244.752	124.428	-1.967	0.049*
R^2 **ajustado**				0.346
$F_{(10,\ 422)}$				23.82

Objetivando-se contornar a existência de variáveis explicativas que não são significativas no modelo de regressão construído, utilizou-se o método de seleção de variáveis *Stepwise*, cujos resultados se apresentam seguidamente (tabela 5.15).

A aplicação do método de seleção de variáveis *Stepwise* permitiu remover as variáveis cuja influência sob a variável produtividade não era significativa, pelo que, de entre as variáveis independentes incluídas inicialmente no modelo, foi removida a variável seleção de varas. Daqui se infere que esta metodologia não influencia significativamente a produtividade das parcelas.

Tabela 5.15: Resultados obtidos por implementação de regressão linear '*lm()*'sem seleção de variáveis pelo método *Stepwise*, a partir do *software* R.

	Estimativa do coeficiente	**Erro Padrão**	***t-value***	***p-value***
Intercept	158.259	24.841	6.371	4.91e−10***
Fr_{nvivas}	102.690	25.124	4.087	5.22e−05 ***
Idade validada	4.512	1.969	2.292	0.022*
Rotação (talhadia)	-66.741	9.653	-6.914	1.75e−11***
AE(2)	-16.922	10.634	-1.591	0.112
Controlo da vegetação(sem controlo)	195.345	102.524	1.905	0.057
RP(3)	66.177	19.419	3.408	0.0007***
RP(4)	25.100	12.502	2.008	0.045*
Entidade (B)	65.516	15.485	4.231	2.86e−05***
Entidade (A)	-194.434	96.678	-2.011	0.045*
R^2 **ajustado**				0.347
$F_{(9,\ 423)}$				26.46

A análise dos resultados obtidos no modelo de regressão supracitado permite evidenciar que os coeficientes de regressão associados às variáveis AE e Controlo de Vegetação não são significativamente diferentes de zero, considerando um nível de significância de 5%. Relativamente à variável controlo de vegetação, sendo o seu *p-value* próximo de 0.05, por forma a sustentar a sua eliminação, apresenta-se na tabela 5.16 os valores de VIF das variáveis do modelo anterior, por forma a avaliar a colinearidade existente entre as variáveis explicativas do referido modelo de regressão.

Tabela 5.16: Resultados da aplicação da função '*VIF()*' do *software* R, ao modelo de regressão com *Stepwise*.

	VIF
FR_{nvivas}	2.069
Idade validada	1.405
Rotação	1.373
AE	1.699
Controlo da vegetação	10.077
RP	1.625
Entidade	13.883

Assim, o valor elevado de VIF inerente à variável controlo da vegetação, conjuntamente com o *p-value* do teste de significância da referida variável, sugerem a sua eliminação. De realçar o valor de VIF elevado relativamente à variável entidade, a qual, porém, sendo significativa no modelo de regressão construído, se opta pela sua não eliminação. Note-se que, anteriormente (secção 5.7.1.1) foi possível detetar a existência de relação entre a idade e a entidade responsável pela parcela, o que pode justificar os valores elevados de VIF inerentes a esta última (tabela 5.16). Porém, tal como explanado, os valores significativos inerentes à entidade responsável pela parcela justificam a sua não remoção no modelo construído subsequentemente.

Desta forma, tendo por base o nível de significância inerente aos referidos coeficientes e o valor de VIF associado à variável controlo da vegetação, é plausível depreender a remoção das variáveis AE e controlo da vegetação. Consequentemente, foi construido de seguida um modelo de regressão, a partir do modelo de regressão obtido com seleção de variáveis *Stepwise*, do qual foram removidas as variáveis AE e controlo da vegetação. Os resultados obtidos pelo modelo de regressão construído sintetizam-se na tabela 5.17.

Tabela 5.17: Resultados obtidos por implementação de regressão linear '*lm()*'com seleção de variáveis usando o método *Stepwise*, a partir do *software* R.

	Estimativa do coeficiente	**Erro Padrão**	***t-value***	***p-value***
Intercept	158.521	20.759	7.636	1.49×10^{-13}***
Fr_{nvivas}	97.417	21.374	4.558	6.77×10^{-6}***
Idade validada	4.056	1.955	2.074	0.039*
Rotação (talhadia)	-72.436	9.390	-7.714	8.77×10^{-14}***
RP(3)	76.839	18.782	4.091	5.14×10^{-5}***
RP(4)	32.235	11.199	2.878	0.0042**
Entidade (B)	67.323	15.258	4.412	$1.30 \times 10{-}5$***
Entidade (A)	-30.556	31.346	-0.975	0.330
		R^2 **ajustado**		0.339
		$F_{(7,425)}$		32.67

A construção deste modelo torna possível a obtenção de coeficientes de regressão significativos.

Assim, com base nos resultados obtidos na construção do último modelo de regressão linear (tabela 5.17), a equação de regressão obtida é:

$$\begin{aligned}\widehat{produtividade} = {} & 158.521 + 97.417 Fr_{nvivas} + 4.056 Idade_{validada} - 72.436 Rotacão(talhadia) \\ & + 76.839 RP(3) + 32.235 RP(4) + 67.323 entd(B) - 30.556 entd(A)\end{aligned} \tag{5.2}$$

Os resultados obtidos permitem deduzir inúmeras conclusões relevantes, mormente no que concerne ao significado dos coeficientes de regressão do modelo, e teste de significância dos mesmos, e ainda em relação ao ajuste do modelo aos dados.

Os coeficientes das variáveis categóricas medem efeitos diferenciais de um grupo relativamente a outro (designado por grupo de referência ou de controlo), isto é, o coeficiente associado à variável qualitativa mede o efeito diferencial do outro grupo relativamente ao grupo de base. Com efeito, é plausível indagar inúmeras conclusões:

- Relativamente ao termo independente estimado no modelo de regressão, o valor obtido instiga que, o valor esperado da produtividade de uma parcela gerida é 158.521 unidades caso se verifiquem valores nulos das variáveis quantitativas inerentes ao modelo, e caso a parcela se encontre em primeira rotação, numa RP classificada por 1, e cuja entidade responsável seja a entidade C. Não obstante, tendo em conta o cariz irrealista de considerar uma parcela com idade nula com produtividade positiva, intui-se que a interpretação inerente a este coeficiente não é pertinente na presente investigação.

- No que concerne à idade das parcelas é plausível intuir que, caso as restantes variáveis consideradas no modelo de regressão se mantenham constantes, em média, a produtividade da parcela aumenta 4.056 unidades, por ano de idade da parcela incrementada.

- No âmbito da variável relativa à frequência relativa de árvores vivas nas parcelas, é verosímil considerar que, por cada unidade desta variável, a produtividade da parcela é incrementada, em média, 97.417 unidades.

- Em comparação com a entidade C, a produtividade aumenta, aproximadamente, 67.323 unidades de medida caso as parcelas sejam geridas pela entidade B, e decresce, aproximadamente, 30.556 unidades de medida, caso as parcelas sejam geridas pela entidade A. As referidas conclusões encontram-se em conformidade com o explanado anteriormente, permitindo evidenciar que a gestão implementada pela entidade B é mais profícua do que a gestão das parcelas implementada pela entidade C, que por sua vez é vantajosa relativamente à gestão desenvolvida pela entidade A.

- A produtividade decresce 72.436 unidades, em média, quando a rotação aplicada é talhadia, em detrimento da aplicação de primeira rotação nas parcelas.

- Comparativamente com parcelas localizadas em regiões de produtividade (RP) classificadas por 2, verifica-se que, em média a produtividade aumenta 76.839 unidades caso as parcelas se localizem em RP classificadas por 3, e aumenta 32.235 unidades caso as parcelas se localizem em RP classificadas por 4. Verifica-se que as conclusões supracitadas vão de encontro ao explanado na secção 4.5, tendo sido assim concluído que, nas regiões estudas, uma RP classificada por 3 permite incrementar a produtividade das parcelas.

Relativamente ao teste de significância global do modelo de regressão construído, dado pelo teste da estatística F, verifica-se que o último modelo de regressão construído permitiu obter um valor de 32.67. O valor crítico da estatística F associado é 2.031. Assim, tendo em conta que o valor de F calculado para o modelo de regressão construído é superior ao valor crítico da estatística F, é plausível concluir a rejeição da hipótese nula, ou seja, rejeita-se que todos os

coeficientes de regressão do modelo são iguais a zero. Consequentemente, o modelo pode ser considerado válido.

Um estudo mais detalhado, relativo ao teste de significância de cada coeficiente, tal como seria expectável, verifica-se que os valores do *p-value* são mais reduzidos (e, consequentemente, mais significativos) após a aplicação do método de seleção de varas *Stepwise*, sendo que tal característica é inerente a todos os coeficientes obtidos (contrariamente ao primeiro modelo de regressão). Com efeito, o *p-value* do teste é bastante reduzido para os oito coeficientes de regressão, o que significa que os valores obtidos para os coeficientes são significativos na população de onde foi retirada a amostra em estudo. Isto é, cada coeficiente é significativamente diferente de zero.

No âmbito do ajuste dos modelos de regressão linear, verifica-se que os valores dos coeficientes de determinação múltiplos ajustados, R^2 são bastante próximos entre si. O último modelo de regressão construído permite obter um valor de R^2 ajustado de 0.339. Com efeito, sendo a diferença entre os valores de R^2 ajustado bastante reduzida, e tendo em conta os testes de significância dos coeficientes do último modelo de regressão construído, este é o modelo de regressão considerado mais adequado, de entre os modelos de regressão estudados. Os valores reduzidos de R^2 ajustado obtidos podem justificar-se na medida em que o modelo de regressão construído usa variáveis explicativas de índole quantitativa e qualitativa, cuja aplicação, de forma geral, implica o declínio do R^2 ajustado. Idealmente, este valor deveria ser próximo de um (o que indicaria um melhor ajuste do modelo de regressão).

Desta forma, tendo em conta que o último modelo de regressão construído (resultados na tabela 5.17) é o que apresenta um melhor ajustamento, efetua-se seguidamente uma análise dos pressupostos do modelo de regressão linear, mormente em relação à colinearidade das variáveis independentes e análise dos resíduos do mesmo. Caso as hipóteses sejam satisfeitas, é plausível concluir que o modelo de regressão se ajusta aos dados, e, consequentemente, faz sentido relacionar as variáveis através da relação linear.

5.8.1 Validação das condições do modelo de regressão linear múltiplo

Seguidamente analisam-se os pressupostos inerentes ao modelo de regressão linear, no que concerne à existência de colinearidade entre as variáveis explicativas, e análise dos resíduos do modelo.

Colinearidade entre as variáveis explicativas

Um dos pressupostos do modelo de regressão linear é a inexistência de colinearidade entre as variáveis independentes. Note-se que o conceito de colinearidade é inerente à existência de correlação entre as variáveis explicativas do modelo, cuja presença pode influenciar de forma negativa os resultados da regressão. Desta forma, numa primeira fase da análise da adequação do modelo de regressão linear construído, é primordial validar este pressuposto. Para tal, fez-se uso da função '*vif()*' do *software* R, cujo valor (*Variance Inflation Factor*) indica o grau de colinearidade entre as variáveis explicativas. A sua aplicação às referidas variáveis do modelo de regressão linear construído, permitiu concluir a inexistência de colinearidade entre as mesmas, na medida em que todos os valores de VIF obtidos são inferiores a 10 (tabela 5.18).

Tabela 5.18: Resultados da aplicação da função '*VIF()*' do *software* R, ao modelo de regressão com *Stepwise* e sem as variáveis controlo da vegetação e área de estudo.

	VIF
FR_{nvivas}	1.481
Idade validada	1.371
Rotação	1.284
RP	1.229
Entidade	1.418

Análise dos resíduos do modelo de regressão

Objetiva-se seguidamente efetuar uma análise dos resíduos do modelo de regressão linear, por forma a verificar as suposições assumidas para os erros inerentes ao modelo de regressão linear construído, e, consequentemente, inferir quanto à adequação do mesmo ao conjunto de dados em estudo. A análise dos resíduos compreende a análise da normalidade, da autocorrelação e da homogeneidade das variâncias dos mesmos.

i) **Normalidade dos resíduos**

Quanto à normalidade dos resíduos, uma análise ao *qqplot* e ao histograma obtidos (figura 5.19), permite evidenciar a existência de apenas um ligeiro afastamento entre os valores reais e os valores teóricos da distribuição normal (*qqplot*) e, no que concerne ao histograma, verifica-se a existência de um pico de frequências seguido de um decrescimento aproximadamente simétrico. Não obstante, os testes de normalidade implementados (tabela 5.19), implicam a conclusão de inexistência de normalidade nos dados.

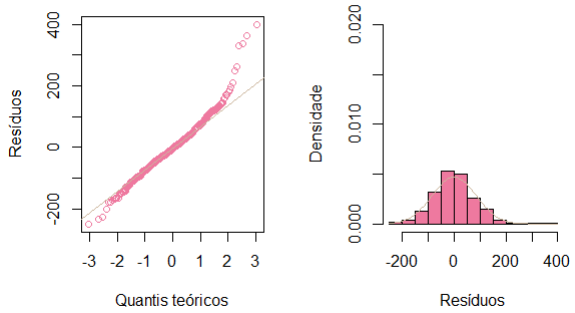


Figura 5.19: *QQplot* e histograma dos resíduos do modelo de regressão linear construído.

Os *p-values* dos testes de hipóteses de Shapiro-Wilk, Anderson-Darling, e Kolmogorov-Smirnov com correção de Lilliefors, são inferiores a 0.05, o que, a um nível de significância de 5%, permite rejeitar a hipótese de Normalidade dos resíduos.

Tabela 5.19: Resultados obtidos por aplicação de testes de Normalidade aos resíduos do modelo de regressão linear construído.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Shapiro-Wilk	2.714e−08	Rejeitar normalidade
Teste de Anderson-Darling	4.360e−05	Rejeitar normalidade
Teste de KS com correção de Lilliefors	0.003	Rejeitar normalidade

A inexistência de conformidade entre as conjeturas indagadas graficamente e a partir de testes de hipóteses, pode justificar-se com base na quantidade de observações em estudo. Com efeito, à medida que a quantidade de observações aumenta, a hipótese nula tem maior probabilidade de ser rejeitada, pois os testes tornam-se mais suscetíveis a pequenos desvios. Assim, visto que os resíduos possuem elevada dimensionalidade, a conclusão da inexistência de normalidade a partir os testes de normalidade não se considera completamente fidedigna neste caso, na medida em que o desvio de observações localizadas na cauda do *qqplot*, produzem influência nos testes de normalidade e, consequentemente, implicam a rejeição da hipótese de normalidade.

ii) **Autocorrelação dos resíduos**

Relativamente à autocorrelação dos resíduos, pressupõe-se que os resíduos do modelo de regressão linear implementado sejam independentes (não correlacionados). A observação da figura 5.20 permite conjeturar a existência de autocorrelação dos resíduos, visível a partir dois grupos de pontos existentes no gráfico, o que evidência que a distribuição dos resíduos em função dos valores preditos não se distribui aleatoriamente em torno da reta horizontal $\varepsilon = 0$.

Assim, por forma a confirmar a presente suposição, foram também implementados três testes: teste de Box-Pierce, teste de Ljung-Box e teste de Durbin Watson, cujos resultados se encontram na tabela 5.20 passíveis de serem consultados. A aplicação concomitante dos três testes permite comprovar as conclusões perquiridas por cada um deles.

Tabela 5.20: Resultados da aplicação de testes à autocorrelação dos resíduos do modelo de regressão linear construído.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Box-Pierce	2.2e−16	Rejeitar autocorrelação nos resíduos
Teste de Ljung-Box	2.2e−16	Rejeitar autocorrelação nos resíduos
Teste de Durbin-Watson	6.664e−08	Rejeitar autocorrelação nos resíduos

Assim, os três testes permitem conjeturar que, tendo sido obtidos valores de *p-value*

reduzidos (inferiores a 0.05), considerando um nível de significância de 5%, é plausível rejeitar a hipótese nula que assume a ausência de autocorrelação dos resíduos.

iii) **Homogeneidade da variância dos resíduos**

Um dos pressupostos inerente aos resíduos do modelo de regressão linear é a homogeneidade das suas variâncias, o qual pode ser testado recorrendo a gráficos e a testes de hipóteses. No que concerne ao gráfico da figura 5.20, é possível detetar a existência de um 'V' localizado horizontalmente, o que permite conjeturar a inexistência de homogeneidade da variância dos resíduos. Tal comprova-se recorrendo a testes de hipóteses de deteção da heteroscedasticidade, em particular, ao teste de Breuch-Pagan e ao teste de Koenker. Tendo em conta que o teste de Breuch-Pagan possui validade assintótica, foi implementado um teste mais robusto: o teste de Koenker.

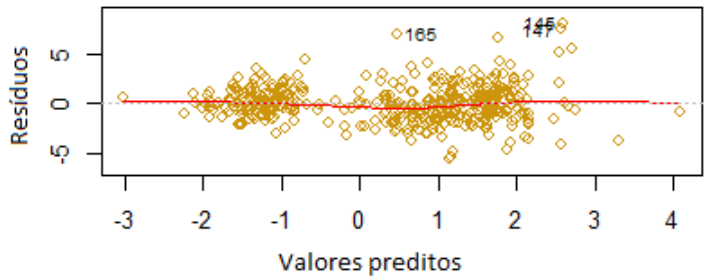


Figura 5.20: Gráfico relativo à heterocedasticidade dos resíduos do modelo de regressão linear construído.

Tabela 5.21: Resultados da aplicação de testes sob a heterocedasticidade dos resíduos do modelo de regressão linear construído. (*Validade assintótica)

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Breusch-Pagan*	3.740×10^{-11}	Rejeitar homogeneidade das variâncias dos resíduos
Teste de Koenker	1.353×10^{-10}	Rejeitar homogeneidade das variâncias dos resíduos

Na sequência do explanado anteriormente, verifica-se que, apesar da inexistência de colinearidade nas variáveis explicativas do modelo de regressão linear construído, os resíduos do modelo não satisfazem os pressupostos inerentes ao modelo de regressão linear. Por conseguinte, é plausível inferir que o modelo de regressão linear construído não se adequa ao conjunto de dados em estudo. Assim, na tentativa de colmatar a referida problemática, ir-se-á seguidamente analisar a caixa de bigodes dos resíduos do modelo de regressão construído, apresentada na figura 5.21, por forma a identificar a presença de *oultliers* nos mesmos e, consequentemente, averiguar a existência de observações influentes.

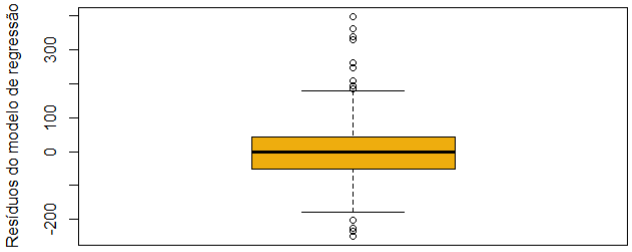


Figura 5.21: Caixa de bigodes dos resíduos do modelo de regressão construído.

No que concerne à caixa de bigodes apresentada, salienta-se primeiramente a simetria inerente à mesma, o que pode indicar a existência de normalidade nos resíduos, o que sustenta a conjetura explorada anteriormente, aquando da análise da normalidade dos resíduos.

A análise da caixa de bigodes dos resíduos do modelo de regressão linear construído permitiu ainda detetar a existência de valores *outliers*. Estes valores correspondem aos resíduos codificados por: 5, 23, 141, 142, 145, 146, 147, 149, 165, 166, 211, 212 e 213.

Assim, com o intuito de avaliar a existência de *outliers* extremos que influenciam severamente os resíduos do modelo de regressão, seguidamente analisa-se a distância de Cook dos resíduos. Para tal, apresenta-se na figura 5.22 o gráfico relativo às distâncias de Cook dos resíduos do modelo, sendo que as observações consideradas influentes se encontram assinaladas.

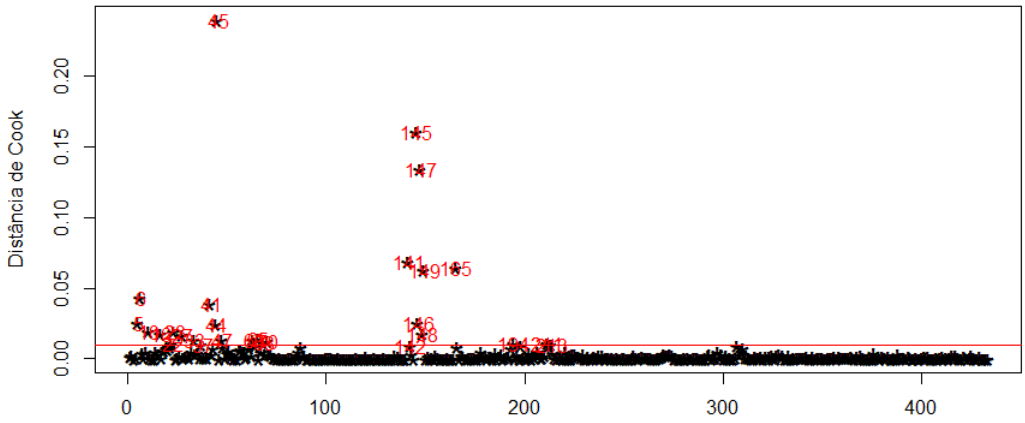


Figura 5.22: Observações influentes relativamente à distância de Cook dos resíduos do modelo de regressão construído.

A figura 5.22 permite detetar a existência de inúmeras observações consideradas influências severas nos resíduos. Com efeito, um estudo comparativo entre os *outliers* dos resíduos detetados a partir da caixa de bigodes e as distâncias de Cook calculadas, permitem intuir que as parcelas cujos resíduos constituem *outliers*, são as parcelas identificadas por: 14, 102, 2423, 2425, 2428, 2429, 2431, 2433, 2565, 2568, 5218, 5219, 5220. Assim, é plausível intuir que, as primeiras duas observações pertencem ao conjunto de dados **Dados inventário 2017**, enquanto que as restantes pertencem ao conjunto de dados inicial **Dados Entidade C**, e constituem valores *outliers* também relativamente às variáveis AMAutil12, Vu12 e S, identificados inicialmente na tabela 1 do Anexo C.1. A remoção das referidas parcelas (simultaneamente ou individualmente) não se mostrou profícua na medida em que não resolve as falhas dos pressupostos de regressão linear múltipla detetados nos resíduos, sendo que estes permanecem não normais, autocorrelacionados e heterocedásticos.

Por outro lado, a construção do gráfico de dispersão entre os valores de produtividade preditos e os respetivos valores reais, que se encontra na figura 5.23, permite sustentar a inadequação de regressão linear múltipla ao conjunto de dados em estudo.

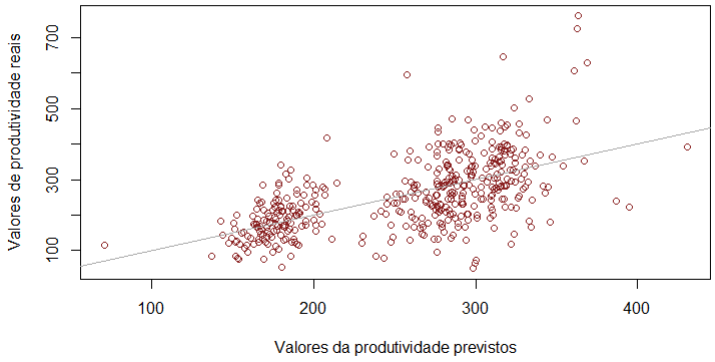


Figura 5.23: Gráfico de dispersão entre os valores de produtividade preditos pelo modelo de regressão construído e os respetivos valores reais.

Com efeito, o afastamento existente entre a bissetriz dos quadrantes ímpares e os pontos referentes à relação entre os valores de produtividade reais e os respetivos valores preditos permitem evidenciar o reduzido valor de R^2 obtido, assim como o reduzido ajustamento do modelo aos dados.

No sentido de tornear a falha dos pressupostos de regressão analisados anteriormente, foram aplicadas metodologias de regressão alternativas, cada vez mais desenvolvidas, nomeadamente, *Support Vector Machine* (SVM), *Neural Networking* (NN) e árvores de regressão. Por forma a selecionar o modelo que melhor se adequa aos dados, a comparação recorreu ao valor de RMSE para os 3 modelos, sintetizados na tabela 5.22.

Tabela 5.22: Comparação de modelos de regressão alternativos aplicados, em termos de valores do RMSE.

	RMSE
***Suport Vector Machine Regression* (SVM)**	81.76
Neural Networking (NN)	271.7
Árvores de Regressão	92.2

Salienta-se, primeiramente, a dimensão dos valores de RMSE, justificada com base nas dimensões inerentes à variável produtividade, cujo valor médio, tal como visto anteriormente, toma o valor 237.42, pelo que, atendendo a literatura consultada, os valores de RMSE obtidos consideraram-se adequados.

Assim, apesar do modelo construído por SVM permitir obter valores de RMSE mais reduzidos, tendo em conta a reduzida interpretabilidade associada ao mesmo, e, com o intuito de contornar as falhas inerentes ao modelo de regressão linear múltipla, aplicou-se regressão com árvores de decisão, cujos resultados se apresentam na secção subsequente.

5.9 Árvores de regressão

Na sequência do desenvolvimento do modelo de regressão linear múltipla, com o intuito de aferir quanto à influência de outros fatores na produtividade das parcelas, e tendo em conta a falha dos pressupostos de regressão linear múltipla inerentes aos resíduos do mesmo, ir-se-á abordar nesta secção um modelo estatístico alternativo: árvores de regressão.

Para a implementação de árvores de regressão, o *software* R possui inúmeros pacotes estatísticos que permitem a sua construção. Entre eles, destacam-se os pacotes: 'rpart', 'tree', 'evtree'. De entre os pacotes mencionados, verificou-se que o que permitiu obter resultados mais aliciantes para o modelo de regressão desenvolvido foi o 'rpart', pelo que ao longo da presente secção serão apresentados os resultados obtidos recorrendo ao mesmo. Salienta-se que a escolha deste pacote na presente investigação prende-se como facto de este ser o que se encontra mais desenvolvido e mais completo no que concerne às técnicas aplicadas.

Assim, com base no estudo desenvolvido na secção anterior, objetiva-se relacionar a variável produtividade (variável dependente) com as variáveis idade da parcela, frequência relativa do número de árvores vivas, rotação, RP e entidade da parcela. Desta forma, as primeiras duas variáveis são de índole quantitativa, enquanto que as restantes são variáveis qualitativas. Pelo facto da variável produtividade ser quantitativa, a árvore construída é uma árvore de regressão (e não de classificação). Note-se que, à semelhança do que foi desenvolvido na secção anterior, também nesta secção foi usado um subconjunto do conjunto de dados, constituído apenas pelas parcelas geridas, por forma a analisar a influência de outros fatores na produtividade das parcelas, e assim intuir quanto à dimensão dessa influência em parcelas geridas, considerando níveis diferenciados de gestão.

Numa primeira fase da aplicação da presente metodologia, foram criados dois subconjuntos a partir dos dados relativos às parcelas geridas: um subconjunto de treino, designado no código implementado por 'dados.tr', usado na construção da árvore de regressão; e um subconjunto de teste, designado por 'dados.te', usado para testar e validar o modelo, isto é, no cálculo dos valores preditos e do valor de RMSE. Assim, na construção do primeiro subconjunto de dados foram usadas 60% das observações do conjunto de dados de parcelas geridas, e para o segundo subconjunto de dados foram usadas as restantes observações. Note-se que a escolha das observações constituintes de cada subconjunto foi aleatória, tendo sido usado o bloco de código subsequente na sua construção.

```
ind=sample(nrow(Parcelas_Geridas), floor(nrow(Parcelas_Geridas)*0.6)) #Indice para selecionar as parcelas

dados.tr=Parcelas_Geridas[ind,] #Dados treino
dados.te=Parcelas_Geridas[-ind,] #Dados teste
```

Seguidamente, fazendo uso dos conjuntos supracitados, foram desenvolvidos dois modelos, que se distinguem pelo uso de *tuning*. Assim, o primeiro modelo é construído recorrendo apenas a re-amostragem e a validação cruzada, enquanto que o segundo recorre ainda a processos de afinação do modelo, designados genericamente por técnicas de *tuning*, que objetivam, idealmente um melhoramento da árvore construída. Ao primeiro ir-se-á designar por árvore 1, e ao segundo por árvore 2. No final, ir-se-á escolher o melhor modelo no que concerne ao valor de RMSE, sendo o modelo escolhido usado em análises subsequentes.

Assim, para a construção do modelo 1, foi usado o código que se encontra no Anexo D.9, o qual permitiu construir a árvore da figura 5.24, antes da poda.

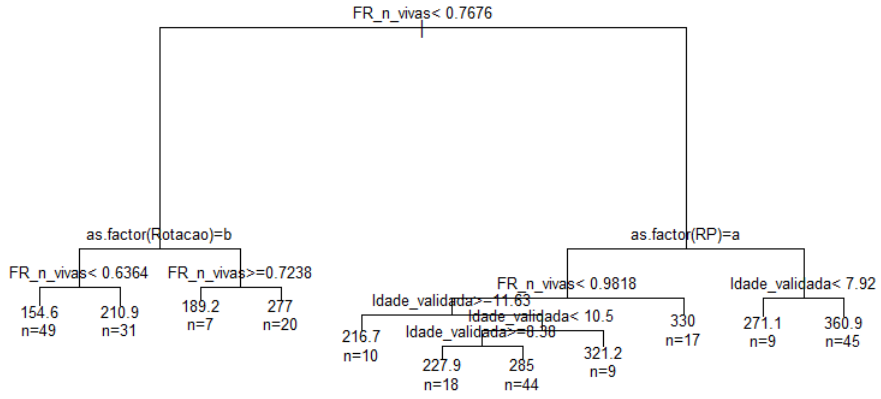


Figura 5.24: Árvore de regressão, segundo modelo de árvore 1, obtida antes da poda.

A árvore de regressão apresentada na figura 5.24 permite intuir que, de entre as variáveis usadas no último modelo de regressão construído, apenas as variáveis FR_{nvivas}, idade da parcela, rotação e RP se encontram na génese da árvore de regressão construída. Desta forma,

de entre as variáveis incluídas no modelo de regressão linear construído, salienta-se que apenas foi removida a variável entidade, o que se coaduna com o *p-value* inerente à mesma no modelo de regressão linear construído.

O pacote 'rpart' possui, intrinsecamente, procedimentos de validação cruzada, a partir dos quais se escolhe o valor ótimo usado na poda da árvore construída, sendo assim a mesma podada recorrendo ao valor calculado. A sua implementação pode ser analisada no bloco de código que se encontra no Anexo D.9 .

A árvore de regressão podada obtida para o presente método pode ser consultada na figura 5.25, tendo permitido a obtenção de um erro de validação cruzada de 82%, aproximadamente, e RMSE de 126.07, aproximadamente. Desta forma, verifica-se que os valores obtidos são relativamente elevados. Salienta-se que os valores obtidos sem efetuar a poda da árvore são superiores, pelo que a aplicação desta metodologia se mostra profícua.

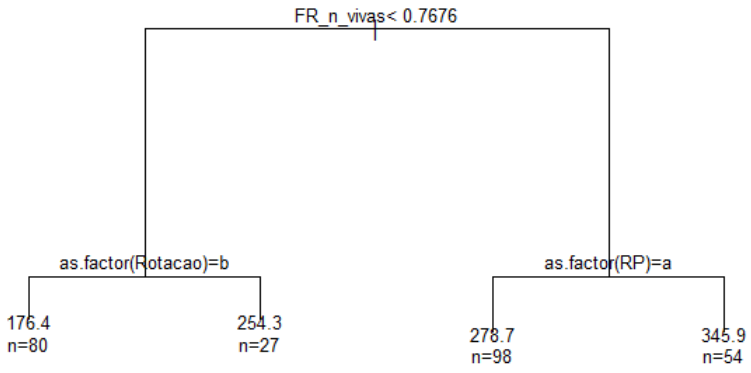


Figura 5.25: Árvore de regressão, segundo modelo de árvore 1, obtida após efetuar a poda.

A aplicação de metodologias de poda na árvore 1, implicou a eliminação da variável idade da parcela, pelo que de entre as variáveis consideradas na primeira árvore construída esta era aquela cujo efeito na produtividade das parcelas era mais reduzido, na árvore de regressão construída.

Seguidamente, considerando aliciante a obtenção de resultados mais robustos, foi implementada a técnica de *tuning* na construção de uma árvore de regressão semelhante à anterior. Tal como referido anteriormente esta árvore foi designada por árvore 2 e o procedimento seguido na sua implementação é apresentado no bloco de código apresentado no Anexo D.9.

Esta técnica de *tunning*, ou seja, de afinamento do modelo de árvores de regressão, baseia-se na construção de uma sequência de valores, no código designado por 'hypergrid', que objetivam adequar os valores inerentes à árvore de regressão o melhor possível, na tentativa de construir uma árvore ótima. Assim, relativamente ao código apresentado, tem-se que os valores de 'hypergrid' são construídos com o intuito de desenvolver uma sequência de números a usar no

parâmetro de controlo da função 'rpart()'. Desta forma são testados inúmeros modelos, com vários valores relativos a estes parâmetros, o que permite alargar a quantidade de modelos em estudo e, consequentemente, escolher o melhor modelo. Para proceder à sua seleção são criadas funções que escolhem, de entre todos os modelos construídos, os valores de erro e de poda que permitem construir o melhor. Após a escolha dos parâmetros adequados, constrói-se o modelo ótimo fazendo uso dos parâmetros que permitem minimizar o erro. A árvore construída sem aplicação de poda encontra-se na figura 5.26.

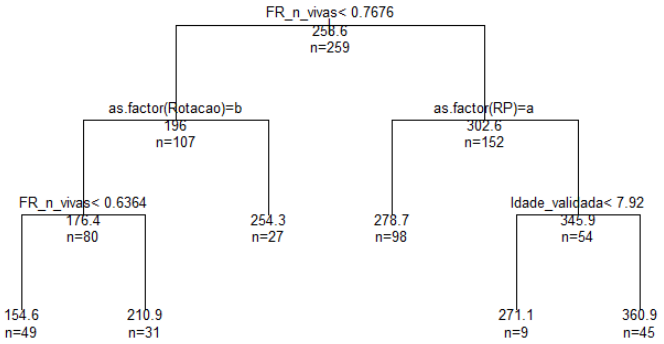


Figura 5.26: Árvore de regressão, segundo modelo de árvore 2, obtida antes de efetuar a poda.

Seguidamente efetua-se a poda da árvore ótima obtida, usando o bloco de código que se apresenta no Anexo D.9.

De forma análoga ao que sucedeu na árvore 1, para a árvore de regressão 2, a aplicação de poda permitiu também a obtenção de valores do erro mais reduzidos. Assim, a aplicação de metodologias de *tuning* e de poda na árvore construída permitiu melhorar os resultados obtidos, sendo que o valor de erro de validação cruzada obtido é de, aproximadamente, 70%, e o RMSE inerente a este modelo é de 92.17, aproximadamente. Consequentemente, a árvore que melhor se adequa ao conjunto de dados é a árvore 2, visto que permitiu obter valores mais reduzidos de erro e de RMSE. Assim, seguidamente são apresentados os resultados obtidos para a árvore ótima.

```
n= 259

node), split, n, deviance, yval
      * denotes terminal node

1) root 259 2883060.0 258.5533
  2) FR_n_vivas< 0.7675817 107  682418.0 196.0456
    4) as.factor(Rotacao)=Talhadia 80  229719.7 176.3992 *
    5) as.factor(Rotacao)=Primeira 27  330327.8 254.2572 *
  3) FR_n_vivas>=0.7675817 152 1488269.0 302.5554
    6) as.factor(RP)=2 98  662859.7 278.6551 *
    7) as.factor(RP)=3,4 54  667835.6 345.9301 *
```

Neste *output*, a primeira linha indica a dimensão da amostra de treino usada, sendo esta constituída por 259 parcelas. Seguidamente, a segunda linha de código informa quanto aos

valores apresentados nas linhas subsequentes, sendo que *node* corresponde ao número do nó, *split* informa quanto à regra usada na divisão, *n* é referente ao número de observações usadas para a divisão, *deviance* é a soma dos quadrados dos resíduos e *yval* designa o valor médio a adotar como resposta (se variável dependente for quantitativa). Note-se que nós que apresentem o símbolo *, correspondem a folhas.

Assim, a raiz da árvore, referente ao primeiro nó, é formada por 259 observações. Este nó é dividido a partir da variável FR_{nvivas}, sendo obtidos dois subgrupos das parcelas: o primeiro, composto por 107 parcelas, onde se verifica a condição $FR_{nvivas} < 0.768$, com *deviance* igual a 682418, aproximadamente, e *yval* na ordem de 196.05; e o segundo subconjunto, composto por 152 parcelas, que verificam a condição $FRnvivas \geqslant 0.768$, com *deviance* e *yval* assumindo os valores 1488269 e 302.56, respetivamente.

No que concerne ao primeiro grupo formado, verifica-se que a divisão recorreu à variável rotação da parcela, tendo sido criados dois subconjuntos: um dos quais é correspondente a parcelas se apresentem em rotação talhadia, composto por 80 parcelas, e com um valor previsto de 176.40 para a produtividade média em parcelas nas condições supracitadas, o qual designa um nó folha; e outro subgrupo de parcelas, designando também um nó folha, caso a rotação das mesmas fosse primeira rotação, formado por 27 parcelas, sendo o correspondente valor de produtividade média estimada de 254.26. Relativamente ao segundo grupo formado, isto é, ao nó 3, este foi dividido com base na variável RP, sendo que no nó 6 pertencem 98 parcelas, cuja região de produtividade (RP) se classifica por 2, e, sendo um nó folha (ou terminal), a produtividade média estimada em parcelas nestas condições é de 278.66. Em relação ao nó 7, este é formado por 54 parcelas, cuja RP se classifica por 3 ou 4, sendo a sua produtividade estimada de 345.93. Desta forma, verifica-se que, nas condições da árvore de regressão construída, as RP 3 e 4 permitem obter valores de produtividade bastante semelhantes entre si.

Seguidamente, na figura 5.27, apresenta-se a representação gráfica da árvore de regressão explanada.

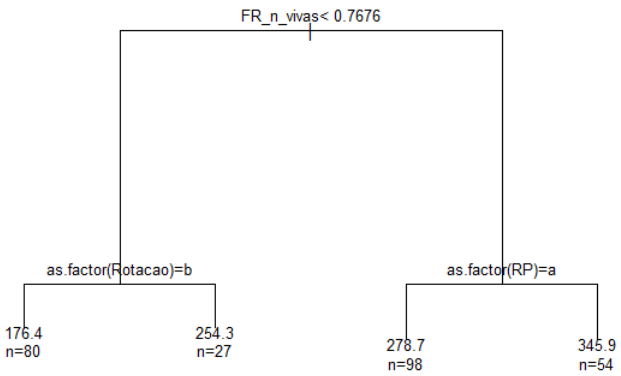


Figura 5.27: Árvore de regressão, segundo modelo de árvore 2, obtida após efetuar a poda.

A árvore de regressão construída permitiu evidenciar que as variáveis que mais influenciam a produtividade de uma parcela são a proporção de árvores que sobreviveram numa parcela (FR_{nvivas}), a rotação e a RP da parcela, visível em ambos os modelos de árvores de regressão construídas.

Assim, a título conclusivo, importa sublinhar que as técnicas de *tuning*, conjugadas com validação cruzada e re-amostragem permitem melhorar o modelo construído. Verifica-se que as árvores de regressão obtidas fazendo uso das referidas metodologias permitem obter árvores iguais, tal como seria expectável. A diferença reside no facto de que, implementando metodologias de afinamento, o número de observações mínimas usadas em cada nó numa divisão antes de ser criado um nó terminal (designado no código por 'minsplit') e o número máximo de nós internos criados entre o nó folha e os nós terminais (designado no código por 'maxdepth') são testados e são escolhidos os melhores de entre a sequência de valores estabelecidos, o que permitem afinar o modelo construído e justificar a obtenção de valores de RMSE distintos, os quais são favoráveis na árvore 2.

A aplicação da presente metodologia permitiu ainda efetuar previsões da produtividade das parcelas, cujos resultados se encontram na tabela 5.23.

Tabela 5.23: Previsão de valores para a produtividade de uma parcela recorrendo à árvore de regressão construída. (*Valor óbtido recorrendo à função '*summary()*' do *software* R aplicada sob a árvore de regressão 2 podada)

		$\widehat{produtividade}$	SSE	MSE*
$FR_{nvivas} < 0.77$	Rotação = talhadia	176.4	229719.7	2871.5
	Rotação=primeira	254.26	330327.8	12234.4
$FR_{nvivas} \geq 0.77$	RP=2	278.66	662859.7	6763.9
	RP=3,4	345.9	667835.6	12367.3

Note-se que as variáveis SSE e MSE, indicam, respetivamente, a soma dos quadrados dos resíduos (dado pelo *deviance* no *output*) e erro quadrático médio (*Mean Squared Error*).

As previsões obtidas para a variável produtividade permitem verificar que a produtividade média assume os valores mais elevados, de 345.9, aproximadamente, quando a proporção de árvores vivas (FR_{nvivas}) é superior ou igual a 0.77, e em regiões de produtividade (RP) classificadas por 3 ou 4, o que vai de encontro às análises implementadas nas secções anteriores. Com efeito, este resultado coaduna-se também com o modelo de regressão linear construído anteriormente, visto que, também neste, um aumento da proporção de árvores sobreviventes na parcela, e uma região de produtividade classificada por 3 ou 4 (em detrimento de RP classificada por 2) permitem incrementar a produtividade. No que concerne ao valor estimado mais reduzido da produtividade, verifica-se que este é observado em parcelas cuja proporção de árvores vivas (FR_{nvivas}) seja inferior a 0.77, e que se encontrem em rotação talhadia. Tal coaduna-se com as conjeturas formuladas na análise de regressão linear da secção anterior, onde

foi possível verificar também que uma diminuição na variável FR_{nvivas}, assim como, parcelas que se localizassem numa rotação do tipo talhadia, comparativamente com parcelas localizadas em primeira rotação, implica a diminuição da produtividade das mesmas.

Por fim, salienta-se a influência das variáveis FR_{nvivas}, rotação e RP na produtividade das parcelas. Esta conclusão é comum à aplicação de ambos os métodos de regressão desenvolvidos: regressão linear e árvores de regressão. Assim, as variáveis referidas anteriormente são as que influenciam mais significativamente a produtividade das parcelas. Neste contexto, um aumento da proporção de árvores sobreviventes na parcela, uma parcela em primeira rotação (comparativamente com parcelas em rotação talhadia) e cuja RP seja classificada por 3 ou 4 (em detrimento de RP classificada por 2) são as condições que permitem maximizar a produtividade.

Capítulo 6

Conclusão e trabalho futuro

O presente relatório de estágio, realizado no âmbito do estágio final do Mestrado em Matemática e Aplicações (no ramo de Estatística e Otimização), teve como principal objeto de estudo, a influência da gestão florestal na produtividade em áreas de minifúndio numa região do centro de Portugal. Assim, para dar resposta ao problema colocado, fez-se uso de dados cujas parcelas se localizam nesta região. Salienta-se o papel preponderante da entidade de acolhimento do estágio curricular na integração da aluna e esclarecimentos dados quando solicitados que ajudaram a tomar consciência da complexidade inerente ao estudo e manuseamento de dados reais, implementando medidas no sentido de resolver problemas característicos dos mesmos.

A seleção de variáveis cuja análise implementada foi mais detalhada em prol da sua importância para o problema em estudo, resultou de uma análise exploratória preliminar efetuada sob o conjunto de dados. Com efeito, a principal característica inerente às mesmas reside na sua influência na produtividade da parcela, e, consequentemente, nos benefícios económicos resultantes da exploração das referidas áreas. A análise exploratória desenvolvida permitiu ainda detetar a existência de cinco grupos de gestão florestal, entre os quais foi comparada a produtividade por forma a intuir quanto à influencia de níveis diferenciados de gestão na produtividade das parcelas. O estudo relativo à influencia da gestão na produtividade das parcelas foi desenvolvido entre os grupos de parcelas com evidencias de gestão e os grupos de parcelas que não apresentam evidencias de gestão.

A Análise em Componentes Principais (ACP) mostrou-se profícua na agregação das seis variáveis referentes à produtividade das parcelas, na medida em que permitiu evitar a escolha de uma só variável, o que implicou a minimização da perda de informação. Assim, o uso de maior quantidade de informação relativa à produtividade da parcela permitiu a obtenção de melhores resultados, mormente no que concerne ao modelo de regressão construído, sendo que os valores de $R^2_{ajustado}$ tomando as variáveis $AMAutil12$ e $Vu12$ como variáveis dependentes (assumindo as mesmas variáveis independentes do último modelo construído) são ambos na ordem de 0.30, aproximadamente. Assim, sendo o valor de $R^2_{ajustado}$ do último modelo construído, aproximadamente, 0.34, verifica-se que este se mostra mais profícuo. De realçar, que nos três modelos construídos (considerando como variáveis dependentes $AMAutil12$, $Vu12$ ou produtividade), verifica-se a falha dos pressupostos inerentes aos resíduos do modelo. Conclui-se assim a adequação da aplicação de ACP aos dados, sendo que a sua aplicação

permitiu explicar cerca de 91% da variabilidade total dos dados. Salienta-se ainda que as restantes metodologias estatísticas desenvolvidas foram aplicadas também ao nível das variáveis originais tendo-se concluído que, à exceção do valor de $R^2_{ajustado}$ mencionado anteriormente, as conclusões indagadas são análogas.

O objetivo inerente à presente investigação incitou o uso de testes paramétricos, tendo por base a consistência e a eficácia inerentes aos mesmos. Não obstante, tendo por base a inadequação dos mesmos ao conjunto de dados, devido à falta de normalidade intrínseca a alguns grupos, foi premente o uso de testes não paramétricos ao longo da presente investigação.

Primeiramente, começou-se por aplicar ANOVA a dois fatores não paramétrico por forma a analisar a influência da gestão e da entidade na produtividade das parcelas e, ainda, a interação existente entre os fatores. Assim, foi possível concluir que ambos os fatores influenciam a produtividade. No entanto, a interação entre ambos é nula. As conclusões deduzidas tornaram possível estudar o efeito de cada um dos fatores na produtividade das parcelas.

Um estudo comparativo entre a produtividade de parcelas geridas e parcelas não geridas permitiu concluir que a aplicação de metodologias de gestão florestal incrementa a produtividade mediana das parcelas, o que, consequentemente, implica que os benefícios económicos associados a essas parcelas são superiores.

No que concerne à influência dos níveis diferenciados de gestão na produtividade das parcelas, com base nos grupos de gestão formados a partir da existência de gestão florestal e das entidades responsáveis pelas parcelas, foi possível concluir que as medianas da produtividade são bastante diferentes para níveis diferenciados de gestão florestal. Neste contexto, a gestão implementada pela entidade B permite a obtenção de resultados de produtividade superiores aos da gestão implementada pela entidade C que, por sua vez, obtém valores de produtividade mais elevados do que a entidade A. Assim, foi possível inferir que, a gestão implementada pela entidade B é a mais rentável na maximização dos benefícios económicos decorrentes da exploração das áreas de minifúndio florestal em estudo.

Tendo consciência que a produtividade da parcela não é apenas influenciada pela gestão florestal e pelas entidades responsáveis pelas parcelas, foi implementada regressão sob as parcelas geridas. Concluiu-se que a produtividade das parcelas é positivamente influenciada pelos fatores: a quantidade de árvores relativas sobreviventes, idade das parcelas, região de produtividade da parcela, e caso a entidade responsável seja a entidade B. Não obstante, verificou-se que a produtividade das parcelas é negativamente influenciada quando a rotação é talhadia, e caso a entidade responsável pelas parcelas seja a entidade A. Visto que os pressupostos inerentes ao modelo de regressão linear não foram todos satisfeitos, alternativamente foi desenvolvida uma metodologia de regressão mais consistente: árvores de regressão, o que possibilitou contornar as falhas supracitadas. A partir da sua implementação foi possível concluir que, de entre as variáveis usadas no último modelo de regressão construído, mormente, a idade das parcelas, a frequência relativa de árvores vivas, rotação, RP e entidade, as variáveis que produzem maior efeito sobre a produtividade das parcelas são a proporção de árvores sobreviventes numa parcela, a rotação e a RP em que se localiza a mesma. Com efeito, analisando o nível de significância associado às variáveis supracitadas no último modelo de regressão linear construído, verifica-se que são as variáveis cujos coeficientes de regressão são mais significativos no modelo. Consequentemente, foi possível concluir que as condições que permitem maximizar a produtividade estimada das parcelas são: a proporção de árvores sobreviventes superior ou

igual a 0.77, e cuja região de produtividade se classifique por 3 ou 4, comparativamente com parcelas em RP 2. .

Sugestões de melhoria e trabalho futuro

O lançamento do inventário florestal recorre a metodologias de Amostragem, efetuando primeiramente a estratificação das áreas de estudo (Amostragem Estratificada), e, após efetuar o cálculo da quantidade de parcelas a lançar em cada estrato (tornando a amostra representativa), faz uso de um programa informático – QGIS - para lançar, de forma aleatória, as coordenadas do número de parcelas previamente calculado (Amostragem Aleatória). Por vezes as coordenadas lançadas aleatoriamente pelo *software* supracitado correspondem a zonas que impossibilitam a recolha ou que influenciam (positiva ou negativamente) as conclusões indagadas a partir da análise dos dados recolhidos. Nessas situações verificou-se que as parcelas eram deslocadas, aquando da recolha dos dados, para parcelas consideradas mais adequadas. Assim, este procedimento decorre do julgamento e da capacidade critica do sujeito responsável pela recolha dos dados, pelo que, a amostragem não é completamente aleatória. De facto, as parcelas são selecionadas de forma intencional pelos sujeitos responsáveis pela recolha dos dados, caso estes considerem que esses elementos possuem características típicas ou representativas da população, o que requer conhecimentos específicos ao nível da população. Com efeito, nestas circunstâncias, a amostra escolhida pode não ser representativa, na sua totalidade, da população. No seguimento do explanado, propõe-se, como possível sugestão de melhoria uma alteração ao nível da metodologia de Amostragem implementada. Com efeito, com o intuito de garantir a representatividade da população que se objetiva estudar, minimizando os erros inerentes e o custo associado a este procedimento, salienta-se primeiramente, a necessidade de reestruturar a estratificação das áreas de estudo. De facto, seria vantajoso efetuar primeiramente um estudo detalhado das áreas de estudo cujos locais não permitem efetuar as medições, mormente, locais de difícil acesso e zonas que incrementam a produtividade (elevada luminosidade, existência de cursos de água, zonas montanhosas,...). Desta forma, o estudo efetuado ao nível dessas zonas permite tomar consciência da sua localização, e, aquando da estratificação das áreas de estudo, permite exclui-las dos estratos construídos. Tal permite que, aquando do lançamento dos pontos aleatórios pelo software QJIS, esses locais que seriam deslocados durante a recolha de dados, não sejam incluídos, minimizando desta forma os erros inerentes ao deslocamento das parcelas, bem como a redução do tempo despendido na recolha de dados visto que não seria gasto tempo no julgamento da adequação das parcelas a amostrar. Consequentemente, é possível reduzir os custos inerentes a este procedimento visto que, sendo o pagamento dos sujeitos efetuado por hora, a redução do tempo de trabalho implica a redução dos custos inerentes ao pagamento supracitado.

Por fim, salienta-se que seria aliciante, em trabalhos futuros, aplicar as sugestões de melhoria supracitadas relativas ao lançamento de amostragem pela reestruturação da estratificação das áreas de estudo, lançando um novo inventário florestal, e, seguidamente, analisar o conjunto de dados resultante, efetuando uma análise comparativa com o conjunto de dados estudado na presente investigação. Tal procedimento seria útil para averiguar a adequação das sugestões de melhoria propostas na minimização dos erros inerentes à recolha de dados.

Tendo em conta a falha dos pressupostos inerentes ao modelo de regressão construído, apesar de estes terem sido colmatados recorrendo a árvores de regressão, seria também aliciante

desenvolver metodologias de regressão alternativas, mormente ANOVA multivariada, a qual, tendo por base as características não paramétricas inerentes ao conjunto de dados, não foi possível desenvolver por falta de tempo útil. Alternativamente, seria também interessante aplicar regressão robusta ao conjunto de dados, a qual, não tendo sido suficientemente desenvolvida, não se mostrou profícua a sua apresentação.

A reflexão crítica inerente à presente investigação permite sublinhar a importância inerente à implementação de uma gestão florestal sustentável, na manutenção da biodiversidade, na diminuição do risco de incêndios florestais, e no incremento da produtividade. Assim, a implementação de técnicas variadas em áreas florestais, como o controlo da vegetação e a reflorestação permitem a maximização dos benefícios económicos resultantes da exploração de áreas de minifúndio florestal. Desta forma, sendo a sociedade moderna cada vez mais caracterizada pelas preocupações não só a nível ambiental, mas também a nível económico, a implementação de práticas de gestão florestal deverão ser cada vez mais uma constante no mundo atual.

Bibliografia

[1] Investigação e Desenvolvimento | The Navigator Company. = http://www.thenavigatorcompany.com/Pasta-e-Papel/Investigacao-e-Desenvolvimento.

[2] RAIZ - Instituto de investigação da Floresta e do Papel | RAIZ. `http://raiz-iifp.pt/`.

[3] Cook's distance in local polynomial regression. *Statistics and Probability Letters 54*, 1 (2001), 33 – 40.

[4] Alboukadel Kassambara. CART Model: Decision Tree Essentials - Articles - STHDA.

[5] Atkinson, A. C. *Plots, transformations, and regression*. Oxford statistical science series. Clarendon Press, Oxford, 1994.

[6] Ben Larson. R: Decision Trees (Regression) – Analytics4All, 2016.

[7] Bohn, F. J., and Huth, A. The importance of forest structure to biodiversity–productivity relationships. *Royal Society Open Science 4*, 1 (2017), 160521.

[8] Breiman, L., Friedman, J., Stone, C., and Olshen, R. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor and Francis, 1984.

[9] Brown, P. J. *Measurement, regression and calibration*. Oxford statistical science series. Clarendon Press, Oxford, 1993.

[10] Cadima, J., and Jolliffe, I. T. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (2016).

[11] Carradori, R. G., and de Siqueira Ramos, P. Avaliação de testes de normalidade implementados no programa R por simulação monte carlo.

[12] Chatterjee, S., and Hadi, A. S. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science 1*, 3 (1986), 379–393.

[13] Clarke, E. M. *Tree-based models*. 1992.

[14] Company, T. N. *Relatório e contas 2015*. 2015.

[15] Conover, W. J. Several k-sample kolmogorov-smirnov tests. *The Annals of Mathematical Statistics 36* (1999).

[16] Cook, R. D. *Detection of Influential Observation in Linear Regression*, vol. 19. [Taylor e Francis, Ltd., American Statistical Association, American Society for Quality], 1977.

[17] Costa, A. R. Instituto RAIZ premiado pela Ordem dos Engenheiros - Vida Rural. = http://www.vidarural.pt/producao/instituto-raiz-premiado-pela-ordem-dos-engenheiros/.

[18] Data Mining. Lecture 10: Regression Trees.

[19] Death, G., and Fabricius, K. E. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology 81*, 11 (2000), 3178–3192.

[20] DeMaris, A. *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Whiley series in probability and statistics. Jonh Wiley and Sons, Inc., Hoboken, New Jersey, 2004.

[21] Dillon, W. R., and Goldstein, M. *Multivariate analysis*. Wiley series in probability and mathematical statistics. John Wiley, New York, 1984.

[22] Dufour, J.-M., Farhat, A., Gardiol, L., and Khalaf, L. Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal 1* (1998), 154–173.

[23] Everitt, B. *An R and S-PLUS companion to multivariate analysis*. Springer texts in statistics. Springer, London, UK, 2005.

[24] Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A. H., and Ståhl, G. Adapting National Forest Inventories to changing requirements - The case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica 48*, 3 (2014), 1–29.

[25] Gama, J., Carvalho, A. P. d. L., Faceli, K., Lorena, A. C., and Oliveira, M. *Extração de conhecimento de dados, Data Mining*, 3 ed. Edições Sílabo, 2017.

[26] Guimarães, C., and Cabral, J. *Estatística*. Mcgraw Hill, Ed., 1997.

[27] Hair Joseph F. JR., Rolph E. Anderson, R. L. T., and Black, W. C. *Multivariate Data Analysis*, 5th edition ed. Prentice-Hall International, Inc., 1998.

[28] Husch, B., Beers, T., and Kershaw, J. *Forest Mensuration*. 4ªedição. Jonh Wiley e Sons, Inc., Hoboken, New Jersey, 2003.

[29] Iles, B. K., and Forestry, J. Some Directions in Forest Inventory.

[30] Johnson, D., Johnson, K., and Hann, D. The importance of forest stand-level inventory to sustain multiple forest values in the presence of endangered species. *Sustainable Forestry: From Monitoring and Modelling to Knowledge Management and Policy Science 97330*, January 2004 (2007), 238–256.

[31] Johnson, R. A. *Wichern, DW (1998), Applied multivariate statistical analysis*, vol. 7632. 81.

[32] Kangas, and Maltamo, M. *Forest Inventory: Methods and Applications*. 2006.

[33] Kim, C., and storer, B. E. Reference values for cook's distance. *Communications in Statistics - Simulation and Computation 25*, 3 (1996), 691–708.

[34] Kim, J.-o., Mueller, C. W., and Lewis-Beck, M. *Factor analysis and related techniques*. 1994.

[35] Kim, Y. J., and Cribbie, R. A. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology 71* (2018), 1–12.

[36] Krzanowski, W. *Recent Advances in Descriptive Multivariate Analysis*. Royal Statistical Society lecture note series. Clarendon Press, Oxford, 1995.

[37] Leotti, V. B., and Rodrigues Birck, A. Comparação dos testes de aderência à normalidade kolmogorov-smirnov, anderson-darling, cramer-von mises e shapiro-wilk por simulação.

[38] Lever, J., Krzywinski, M., and Altman, N. Principal component analysis. *Nature Publishing Group 14*, 7 (2017).

[39] Lewis-Beck, M. S. *Regression analysis*. International handbooks of quantitative applications in the social sciences. Sage, London, 1993.

[40] Li, X., Qiu, W., Morrow, J., Demeo, D. L., Weiss, S. T., Fu, Y., and Wang, X. A Comparative Study of Tests for Homogeneity of Variances with Application to DNA Methylation Data.

[41] Lim, T. S., and Loh, W.-Y. A comparison of tests of equality of variances. 287–301.

[42] Lim, T.-S., and Loh, W.-Y. A comparison of tests of equality of variances. *Computational Statistics and Data Analysis 22*, 3 (1996), 287 – 301.

[43] Loh, W.-Y. Classification and regression trees. *John Wiley and Sons, Inc 1* (2011).

[44] Lopes, M. D. M., Castelo Branco, V. T. F., and Soares, J. B. Utilização dos testes estatísticos de Kolmogorov-Smirnov e Shapiro-Wilk para verificação da normalidade para materiais de pavimentação. *Transportes 21*, 1 (2013), 59.

[45] M. Mendes de Oliveira, L. D. S., and Fortuna, N. *Econometria*, 1a edição ed. Escolar Editora, 2011.

[46] Mahibbur Rahman, M., and Govindarajulu, Z. A modification of the test of shapiro and wilk for normality. 219–236.

[47] Maindonald, J., and Braun, J. *Data Analysis and Graphics using R - an Example-based Approach*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, United States of America, 2003.

[48] Manly, B. F., and Alberto, J. A. N. *Multivariate statistical methods: a primer*. CRC Press, 2016.

[49] Manoukian, E. B. *Mathematical nonparametric statistics*. Gordon and Breach Science Publishers, Inc., 1986.

[50] Mardia, K., Kent, J., and Bibby, J. *Multivariate analysis*. Probability and mathematical statistics. Academic Press, London, 1994.

[51] Marques, C. P., Fonseca, T. F., and Duarte, J. C. *Guia prático de avaliações florestais - Dendrometria*. 1ªedição. Sílabas e Desafios, 2017.

[52] Marôco, J. *Análise Estatística com o PASW Statistics*, 3a edição ed. Sílabo, Ed., 2010.

[53] Mati, A. & Dawaki, S. A. Role of Forest Inventory in Sustainable Forest Management : A Review. 33–40.

[54] Moessner, K. E. Photo interpretation in forest inventories. *Photogr. Engin.* (1953), 496–507.

[55] Mohd Razali, N., and Yap, B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests.

[56] Moisen, G. G. Classification and regression trees. *US Forest Service* (2008).

[57] Mood, A., Graybill, F., and Boes, D. *Introduction to the Theory of Statistics*. International Student edition. McGraw-Hill, 1974.

[58] Morgan, J. Classification and Regression Tree Analysis.

[59] Murteira, B., C.S.Ribeiro, e Silva, J. A., and Pimenta, C. *Introdução à Estatística*. Escolar Editora,Lisboa, 2010.

[60] Nelson, M., Moisen, G., Finco, M., and Brewer, K. Forest inventory and analysis in the United States: Remote sensing and geospatial activities. *Photogrammetric Engineering and Remote Sensing 73*, 7 (2007), 729–732.

[61] Nelson, M. D., McRoberts, R. E., and Hansen, M. C. Forest land area estimates from Vegetation Continuous Fields. *Remote Sensing for Field Users: Proceedings of the Tenth Biennial Forest Service Remote Sensing Applications Conference* (2004), 5–9.

[62] Powers, D. A., and Xie, Y. *Statistical Methods for Categorical Data Analysis*. Academic Press, 2000.

[63] RAIZ. Centro de Investigação da Floresta e Papel | RAIZ. = http://raiz-iifp.pt/instituto/.

[64] Razali, N. M., and Wah, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics 2* (2011), 21–33.

[65] Sawilowsky, S. S. Nonparametric tests of interaction in experimental design. *Review of Educational Research 60*, 1 (1990), 91–126.

[66] Scott, C. T., Gove, J. H., El-shaarawi, A. H., and Piegorsch, W. W. Forest inventory. 814–820.

[67] Shaw, J. Benefits of a strategic national forest inventory to science and society: the USDA Forest Service Forest Inventory and Analysis program. *iForest - Biogeosciences and Forestry 1*, 1 (2008), 81–85.

[68] Siegel, S. *Nonparametric statistics for the behavioral sciences.* McGraw-Hill series in psychology. McGraw-Hill Kogakusha, Tokyo, 1956.

[69] Soave, D., and Sun, L. A generalized levene; scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics 73*, 3 (9 2017), 960–971.

[70] Sonnberger Harald, D. A. Belsley, K. K., and Welsch, R. E. Regression diagnostics - identifying influential data and sources of collinearity. *Journal of Applied Econometrics 4*, 1 (1980), 97–99.

[71] Soporcel, G. P. Investigação nas áreas da floresta e do papel. uma renovação de raiz. 1–5.

[72] Stapleton, J. H. *Linear statistical models*, vol. 719. John Wiley e Sons, 2009.

[73] Stapleton, J. H. *Models for probability and statistical inference.* Wiley series in probability and statistics. Wiley Interscience, Hoboken (NJ), cop. 2008.

[74] Tewari, V. P. Forest inventory, assessment, and monitoring, and long-term forest observational studies, with special reference to India. *Forest Science and Technology 12*, 1 (2016), 24–32.

[75] Torman, V. B. L., Coster, R., and Riboldi, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista HCPA 32*, 2 (2012), 227–234.

[76] Torrão, S. Os Eucaliptos e as Aves da Quinta de São Francisco.

[77] Winters, C. A., Moore, C. F., Kuntz, S. W., Weinert, C., Hernandez, T., and Black, B. Principal components analysis to identify influences on research communication and engagement during an environmental disaster. *BMJ Open 6*, 8 (2016).

[78] Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems 2*, 1 (1987), 37 – 52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[79] Yamashita, T., Yamashita, K., and Kamimura, R. A stepwise aic method for variable selection in linear regression. *Communications in Statistics - Theory and Methods 36*, 13 (2007), 2395–2403.

[80] Young, M. H. P. *Statistical Analysis For Decision Making*, 6th edition ed. Liz Widdicombe, 1993.

Anexo A

Complementos ao capítulo 2 - O estágio curricular

A.1 Lançamento do inventário florestal

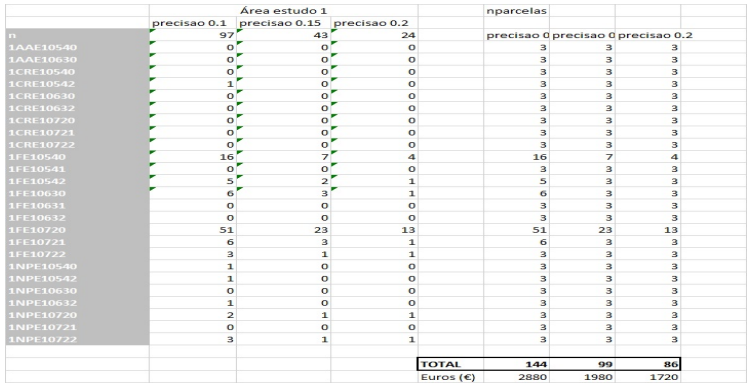
	Área de Estudo 1																														
	1AAE10540	1AAE10630	1CRE10540	1CRE10542	1CRE10630	1CRE10632	1CRE10720	1CRE10721	1CRE10722	1FE10540	1FE10541	1FE10542	1FE10630	1FE10631	1FE10632	1FE10720	1FE10721	1FE10722	1NPE10540	1NPE10542	1NPE10630	1NPE10632	1NPE10720	1NPE10721	1NPE10722	TOTAL					
ÁREA	72872,41821	73571,8239	13562,59659	135159,1609	15424,70622	548,433131	17724,50135	20171,61378	54398,21727	3116133,6	9570,632767	962741,1236	1087418,945	1961,2427	10802,42	9575630,259	1116086,98	586171,622	193036,688	108007,3036	32264,67528	188271,3669	391382,4801	10830,35048	545780,0891	18343523,25					
PARCELA	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400					
VOLUME MEDIO	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607					
DESVIO-PADRAO	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223					
N	182	184	34	338	49	1	44	50	136	7790	24	2407	2719	5	27	23939	2790	1465	483	270	81	471	978	27	1364	45859					
PRECISAO																										0,1	0,15	0,2			
E																										16	23	31			
CUSTO																										20					

(a)

	Área de Estudo 2																			
	2CRE9720	2CRE9721	2FE10540	2FE10541	2FE10720	2FE10721	2FE10722	2FE8730	2FE8731	2FE9720	2FE9721	2FE9722	2NPE10720	2NPE10721	2NPE10722	2NPE9720	2NPE9721	2NPE9722	TOTAL			
ÁREA	123134,6331	1143,9749	1192791,055	5957,889319	5343044,213	492893,7106	479989,0834	48205,55447	13,696615	14385449,12	2086719,081	688621,9935	213598,0984	67798,409	727,4419	938002,502	462980,368	12560,2983	26543631,13			
PARCELA	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400			
VOLUME MEDIO	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607	155,607			
DESVIO-PADRAO	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223	76,6223			
N	308	3	2982	15	13358	1232	1200	121	0	35964	5217	1722	534	169	2	2345	1157	31	66359			
PRECISAO																			0,1	0,15	0,2	
E																			16	23	31	
CUSTO																			20			

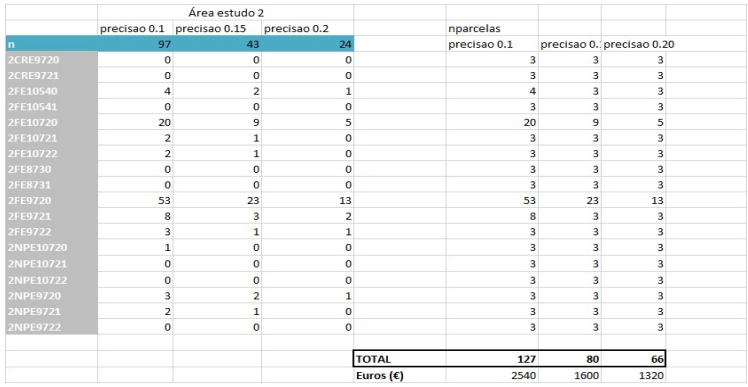
(b)

Figura 1: Cálculos efetuados para o lançamento do inventário florestal, com o auxílio da página de cálculo excel, para cada área de estudo. (a) Área de estudo 1. (b) Área de estudo 2.



	Área estudo 1				nparcelas		
	precisao 0.1	precisao 0.15	precisao 0.2				
n	97	43	24		precisao 0	precisao 0	precisao 0.2
1AAE10540	0	0	0		3	3	3
1AAE10630	0	0	0		3	3	3
1CRE10540	0	0	0		3	3	3
1CRE10542	1	0	0		3	3	3
1CRE10630	0	0	0		3	3	3
1CRE10632	0	0	0		3	3	3
1CRE10720	0	0	0		3	3	3
1CRE10721	0	0	0		3	3	3
1CRE10722	0	0	0		3	3	3
1FE10540	16	7	4		16	7	4
1FE10541	0	0	0		3	3	3
1FE10542	5	2	1		5	3	3
1FE10630	6	3	1		6	3	3
1FE10631	0	0	0		3	3	3
1FE10632	0	0	0		3	3	3
1FE10720	51	23	13		51	23	13
1FE10721	6	3	1		6	3	3
1FE10722	3	1	1		3	3	3
1NPE10540	1	0	0		3	3	3
1NPE10542	1	0	0		3	3	3
1NPE10630	0	0	0		3	3	3
1NPE10632	1	0	0		3	3	3
1NPE10720	2	1	1		3	3	3
1NPE10721	0	0	0		3	3	3
1NPE10722	3	1	1		3	3	3
				TOTAL	144	99	86
				Euros (€)	2880	1980	1720

(a)



	Área estudo 2				nparcelas		
	precisao 0.1	precisao 0.15	precisao 0.2				
n	97	43	24		precisao 0.1	precisao 0.	precisao 0.20
2CRE9720	0	0	0		3	3	3
2CRE9721	0	0	0		3	3	3
2FE10540	4	2	1		4	3	3
2FE10541	0	0	0		3	3	3
2FE10720	20	9	5		20	9	5
2FE10721	2	1	0		3	3	3
2FE10722	2	1	0		3	3	3
2FE8730	0	0	0		3	3	3
2FE8731	0	0	0		3	3	3
2FE9720	53	23	13		53	23	13
2FE9721	8	3	2		8	3	3
2FE9722	3	1	1		3	3	3
2NPE10720	1	0	0		3	3	3
2NPE10721	0	0	0		3	3	3
2NPE10722	0	0	0		3	3	3
2NPE9720	3	2	1		3	3	3
2NPE9721	2	1	0		3	3	3
2NPE9722	0	0	0		3	3	3
				TOTAL	127	80	66
				Euros (€)	2540	1600	1320

(b)

Figura 2: Cálculo do número n de parcelas a amostrar em cada área de estudo com o auxílio da página de cálculo excel. (a) Área de estudo 1. (b) Área de estudo 2.

Anexo B

Complementos ao capítulo 3 - Revisão da metodologia estatística

B.1 Tabela relativa à ANOVA a dois fatores

Tabela 1: Organização da variável dependente Y, em função dos dois fatores em estudo A e B, para a aplicação de ANOVA a dois fatores.

		Fator B			
		1	**2**	**...**	**b**
Fator A	**1**	y_{111}	y_{121}	$\cdots$	y_{1b1}
		y_{112}	y_{122}	$\cdots$	y_{1b2}
		...	...	...	...
		y_{11r}	y_{12r}	$\cdots$	y_{1br}
	2	y_{211}	y_{221}	$\cdots$	y_{2b1}
		y_{212}	y_{222}	$\cdots$	y_{2b2}
		...	...	...	...
		y_{21r}	y_{22r}	$\cdots$	y_{2br}
	...	...	...	...	...
	a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}
		y_{a12}	y_{a22}	$\cdots$	y_{ab2}
		...	...	...	...
		y_{a1r}	y_{a2r}	$\cdots$	y_{abr}

B.2 Representações gráficas que exemplificam a existência e inexistência de homogeneidade de variâncias.

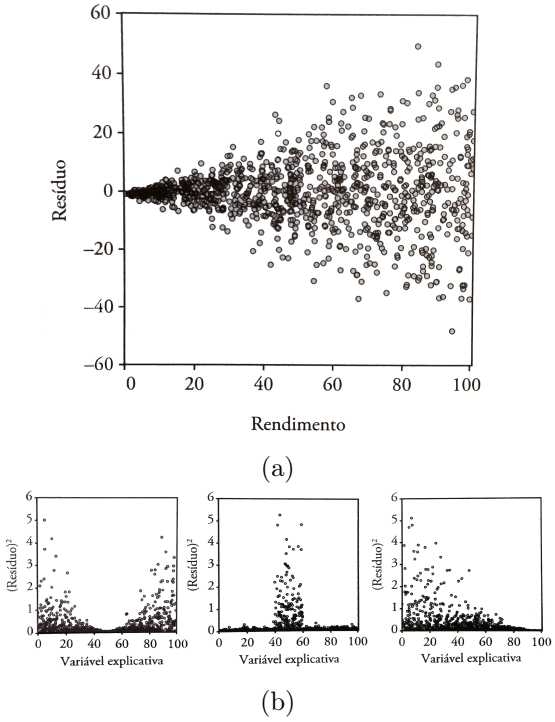


Figura 1: Ambas as figuras exemplificam a existência de heterocedasticidade das variâncias dos dados.

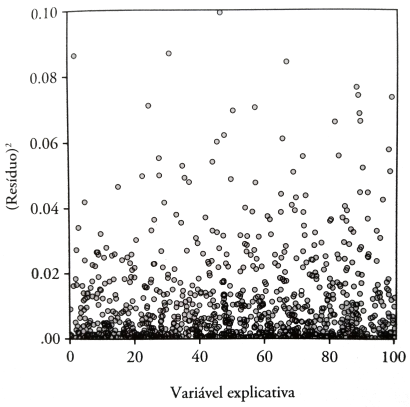


Figura 2: Exemplifica a existência de homogeneidade das variâncias dos dados (inexistência de heterocedasticidade das variâncias dos dados.

Anexo C

Complementos ao capítulo 4 - O problema e os dados

C.1 Deteção de *outliers* no conjunto de dados relativos à entidade C

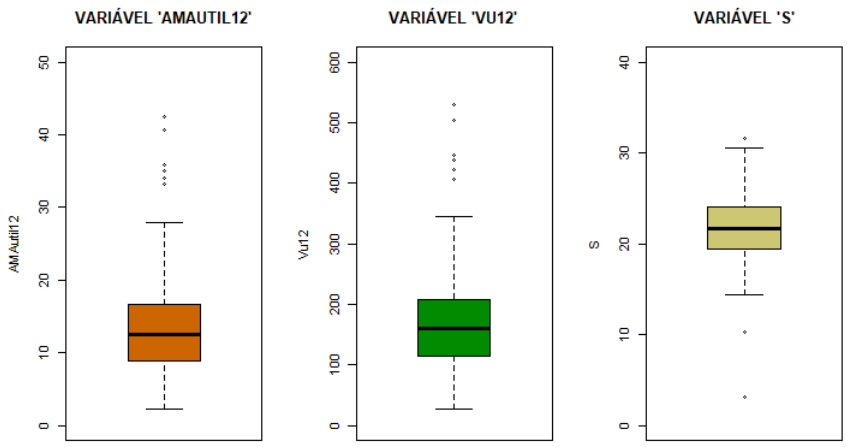


Figura 1: Caixas de bigodes relativas às variáveis AMAutil12, Vu12 e S para deteção de *outliers*.

Visualmente, os *outliers* são representados pelos pontos representados nas caixas de bigodes. A identificação das parcelas correspondentes a esses pontos sintetiza-se na tabela 1, cuja eliminação não alterou os resultados obtidos.

Tabela 1: Identificação das parcelas *outliers* pertencentes à entidade C, relativamente às variáveis AMAutil12, Vu12 e S.

	AMAutil12	**Vu12**	**S**
Parcelas *outliers*	2423; 2428; 2429; 2431; 2433; 2565.	2423; 2428; 2429; 2431; 2433; 2565.	2425; 4860; 5218; 5219; 5220.

C.2 Análise exploratória da variável FR_{nvivas}

O gráfico divergente entre os valores médios da variável FR_{nvivas} e cada grupo de gestão, que se encontra na figura 2 permite intuir que, em média, a quantidade de árvores vivas é superior em parcelas que apresentam evidencias de gestão, cuja entidade responsável é a B. Por outro lado, é possível intuir que a quantidade de árvores que sobrevivem é, em média, inferior em parcelas geridas pela entidade C.

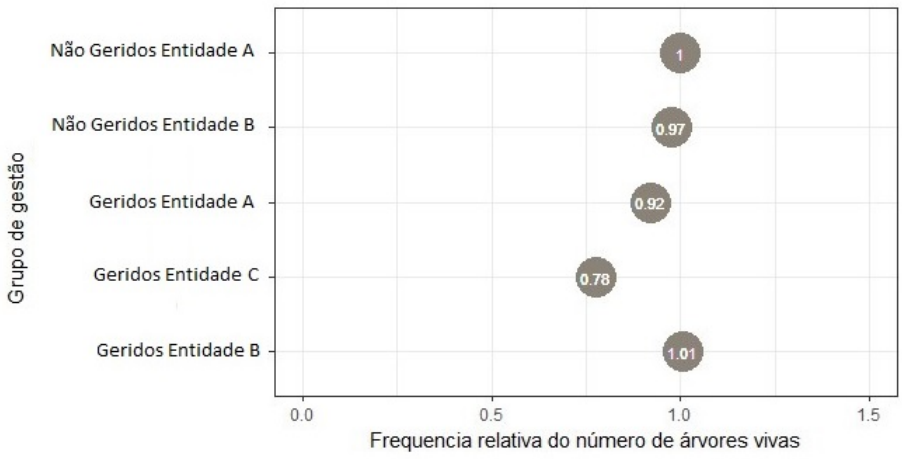


Figura 2: Gráfico divergente dos valores médios da variável FR_{nvivas} por grupo de gestão.

A análise do gráfico apresentado seguidamente encontra-se em conformidade com as conclusões supracitadas na medida em que, a maioria das parcelas geridas pela entidade C (representadas pela mancha verde) possuem valores de frequência relativa de árvores vivas inferiores a 1. As parcelas nestas condições apresentam valores de AMAutil12 mais elevados do que os restantes grupos em estudo.

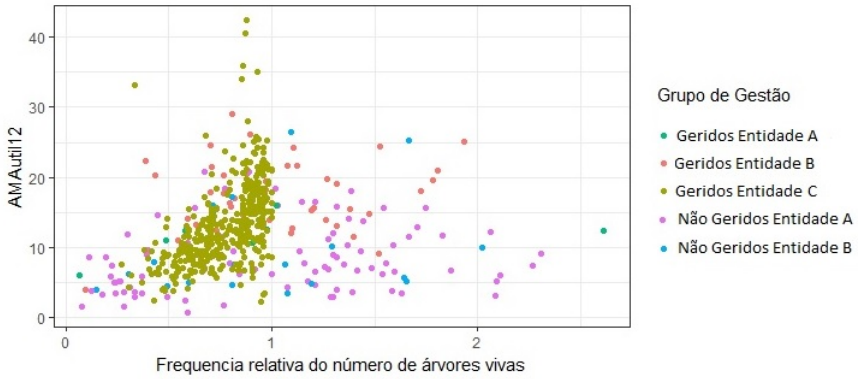


Figura 3: Gráfico de dispersão entre o AMAutil12 e a variável FR_{nvivas} por grupo de gestão.

C.3 Características sumárias das variáveis em estudo

Seguidamente apresentam-se as características sumárias de variáveis consideradas pertinentes nos estudos e análises elaborados.

a) **Variáveis quantitativas:**

Tabela 2: Características sumárias de algumas variáveis quantitativas de relevo.

Variável	Média	Desvio Padrão	1º Quartil	Mediana	3º Quartil	Mínimo	Máximo
AMAutil1	12.677	6.151	8.30	11.900	16.25	0.650	42.400
dg12	13.625	3.287	11.50	13.500	15.36	6.200	38.750
FR árvores vivas	0.843	0.338	0.656	0.845	0.941	0.069	2.611
G	14.246	8.107	8.065	13.440	19.445	0.250	52.990
G12	20.051	7.260	15.18	19.610	24.57	2.710	55.530
Hdom	17.630	5.600	13.20	17.630	22.10	4.000	32.200
Hdom12	23.018	4.290	20.65	23.300	25.90	4.200	33.600
Idade	7.906	2.712	5.580	8.340	10.510	2	14
N	1169.525	511.281	824.12	1099	1423	49.950	3396
NPL	1420.620	404.935	1149.7	1423	1623.0	299.679	4045.666
S	20.924	4.221	18.53	21.140	23.78	3.210	31.50
V	117.988	89.971	47.60	101.846	168.17	0.720	3396
V12	194.306	90.788	128.83	184.328	247.22	12.750	632.722
Vu	97.232	75.108	38.45	83.883	138.77	0.560	478.823
Vu12	161.171	76.298	106.25	153.044	205.78	10.230	529.868
Vmdi	106.275	88.042	36.38	91.799	153.67	0.030	552.742
Vmudi	87.811	73.612	29.27	75.150	127.64	0.020	462.941
Vmudi12	152.132	73.812	99.31	143.385	194.85	7.800	508.992

C.4 Variáveis qualitativas:

Tabela 3: Relação entre valores médios da variável *Site Index*, S, relativamente às variáveis gestão, região de produtividade e rotação das parcelas.

Variáveis	Número de níveis	Níveis	Nºde parcelas por nível
Seleção de Variáveis	2	Sim	434
		Não	109
Preparação Terreno	7	Cava	26
		Desalinhado	61
		Linha	25
		Ripagem	32
		Sem armação	170
		Terraços	136
		Vala e Cômoro	93
Controlo Vegetação	2	Com controlo	453
		Sem controlo	90
Rotação	3	Misto	2
		Primeira	259
		Talhadia	282
Área de estudo	2	1	300
		2	243
Clima	3	8	5
		9	113
		10	425
Solo	3	5	114
		6	35
		7	394
RP	3	2	389
		3	40
		4	114

C.5 Análise da variável S, *Site Index*

Tabela 4: Relação entre valores médios da variável *Site Index*, S, relativamente às variáveis gestão, região de produtividade e rotação das parcelas.

		RP2	**RP3**	**RP4**
Com Gestão	**1ªrotação**	22.03	26.68	24.69
	Talhadia	20.39	20.08	17.99
Sem Gestão	**1ªrotação**	20.06	17.08	16.43
	Talhadia	17.46	16.25	17.21

C.6 Contabilização do número de parcelas em cada grupo de gestão formado

Tabela 5: Contabilização do número de parcelas existentes em cada entidade, geridas e não geridas em frequência absoluta.

	Não Gerido	**Gerido**
Entidade A	91	8
Entidade B	19	48
Entidade C	0	377

Tabela 6: Contabilização do número de parcelas existentes em cada entidade, geridas e não geridas em percentagem.

	Não Gerido (%)	**Gerido** (%)
Entidade A	91.919	8.081
Entidade B	28.358	71.642
Entidade C	0.000	100.000

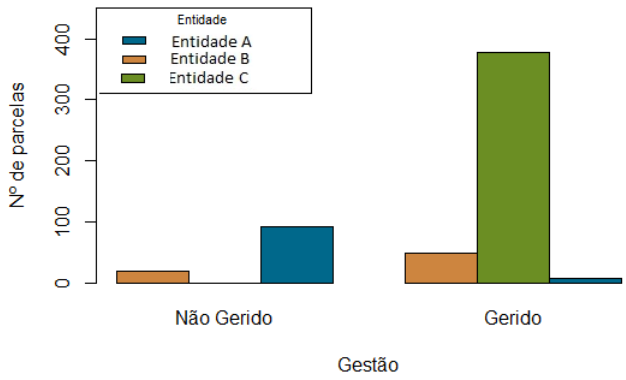


Figura 4: Gráfico de barras que contabiliza o número de parcelas pertencentes a cada entidade, consoante se verifique ou não aplicação de técnicas de gestão florestal.

C.7 Contabilização do número de parcelas, por grupo de gestão, em cada área de estudo.

Na análise subsequente objetiva-se intuir quanto à quantidade de parcelas de cada grupo de gestão estabelecido na secção 4.4.2 em cada área de estudo, verificando-se a equidade na distribuição dos grupos de gestão em cada área de estudo.

Tabela 7: Contabilização das parcelas em cada área de estudo, por grupo de gestão, em frequência absoluta.

	AE1	**AE2**
Geridos Entidade A	1	7
Geridos Entidade B	24	24
Geridos Entidade C	203	174
Não Geridos Entidade A	60	31
Não Geridos Entidade B	12	7

Tabela 8: Contabilização das parcelas em cada área de estudo, por grupo de gestão, em percentagem.

	AE1 (%)	**AE2** (%)
Geridos Entidade A	12.500	87.500
Geridos Entidade B	50.000	50.000
Geridos Entidade C	53.846	46.154
Não Geridos Entidade A	65.934	34.066
Não Geridos Entidade B	63.158	36.842

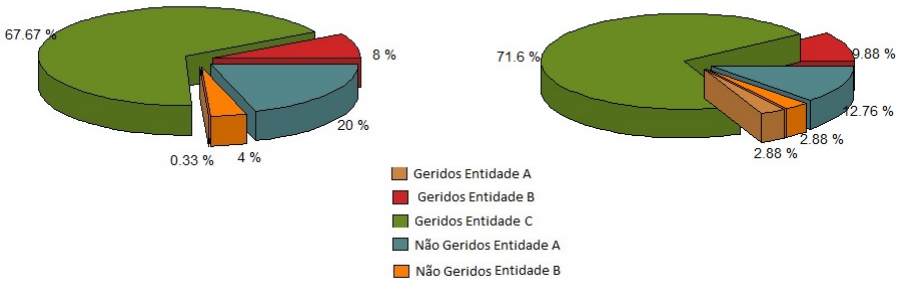


Figura 5: Gráficos circulares relativos à contabilização das parcelas de cada grupo de gestão, por área de estudo, sendo à esquerda o gráfico relativo à área de estudo 1, e à direita o gráfico relativo à área de estudo 2.

C.8 Relação entre RP das parcelas e a sua produtividade

As condições do clima e do solo, influenciam substancialmente a produtividade da parcela. Assim, sendo RP a variável que conjuga as informações das variáveis supracitadas, seguidamente analisa-se de forma detalhada essa influência. Note-se que no presente anexo contabiliza-se a produtividade das parcelas recorrendo aos valores do AMAutil12, visto ser a variável de maior interesse para a organização de acolhimento nesse sentido.

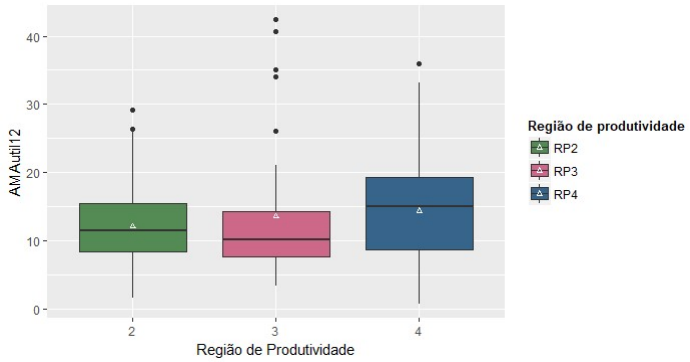


Figura 6: Caixas de bigodes da RP em função do AMAutil12.

Tendo por base as caixas de bigodes apresentadas na figura 6 verifica-se que a RP4 permite incrementar a produtividade das parcelas, enquanto que a RP2 minimiza os valores médios da referida variável. As conclusões indagadas são considerados contraditórios tendo por base a definição da variável RP, na medida em que, por definição, quanto maior o valor de RP menos favorável à Produtividade é essa região. A contradição dos resultados obtidos relativamente ao que seria expectável justifica-se tendo por base erros inerentes à recolha dos dados.

A figura 7 permite caracterizar de forma detalhada cada área de estudo, diferenciada por região de produtividade, em função do valor de AMAutil12 da parcela, para cada área de estudo. De uma forma geral é plausível constatar que na área de estudo 1, as parcelas possuem valores de AMAutil12 médios superiores aos valores médios existentes em parcelas pertencentes à área de estudo 2, para todos os níveis de RP. Relativamente às parcelas localizadas na área de estudo 1, constata-se que a região de produtividade onde o AMAutil12 médio é incrementado é a RP4, apesar de as três permitirem obter valores de produtividade média bastante próximos entre si. Em parcelas localizadas na área de estudo 2, verifica-se que o AMAutil12 médio é minimizado quando a região de produtividade se classifica por 4, sendo maximizada quando a região de produtividade se classifica com 3.

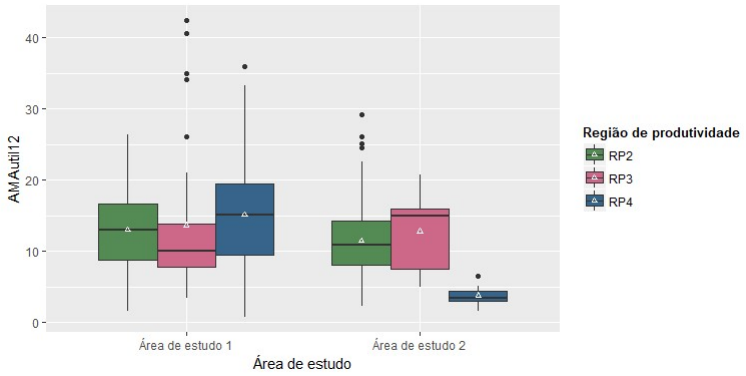


Figura 7: Caixas de bigodes entre cada área de estudo e a produtividade das parcelas, em função da RP das mesmas.

Anexo D

Complementos ao capítulo 5 - Análise dos dados

D.1 Gráficos de dispersão entre as variáveis usadas na ACP.

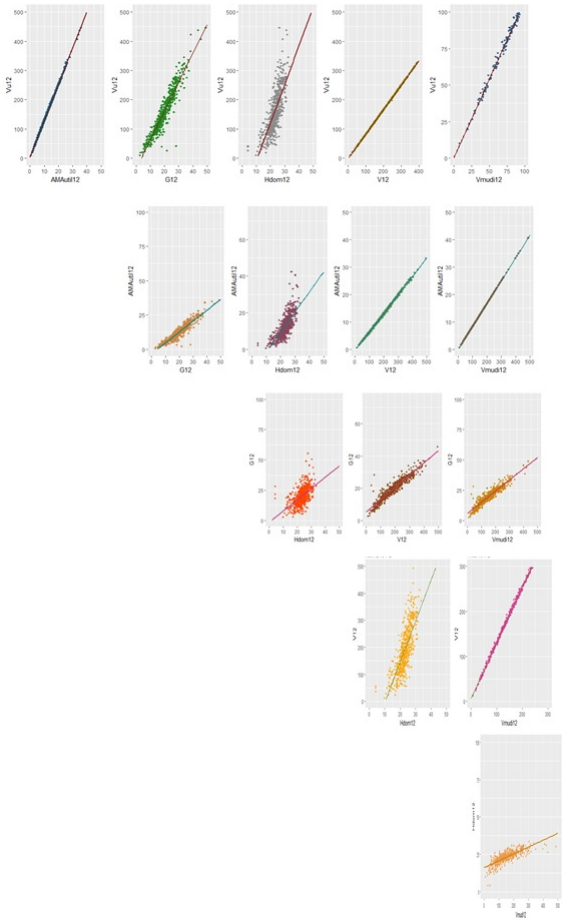


Figura 1: Gráficos de dispersão entre as variáveis usadas na ACP.

D.2 Análise da adequação dos testes paramétricos aos grupos de parcelas geridas e não geridas relativamente à variável produtividade.

A análise da adequação dos testes paramétricos na comparação da produtividade entre parcelas geridas e não geridas, pressupõe a existência de Normalidade e a homogeneidade das variâncias relativamente a esta variável nos grupos em estudo.

Normalidade dos grupos de parcelas Geridas e Não Geridas

A análise da Normalidade nos grupos de parcelas que apresentam gestão e que não apresentam gestão, faz uso primeiramente de uma análise gráfica, por forma a formular conjeturas quanto à Normalidade dos grupos referidos, e seguidamente comprova-se as conclusões formuladas utilizando testes de Normalidade.

A análise gráfica inicia-se com a construção de um gráfico que relaciona a função densidade de probabilidade estimada relativa à variável produtividade entre os dois grupos em estudo. Com efeito, a sua construção torna possível a analise da existência de semelhanças entre as funções densidades de probabilidade estimadas construídas e a função densidade de probabilidade da distribuição Normal.

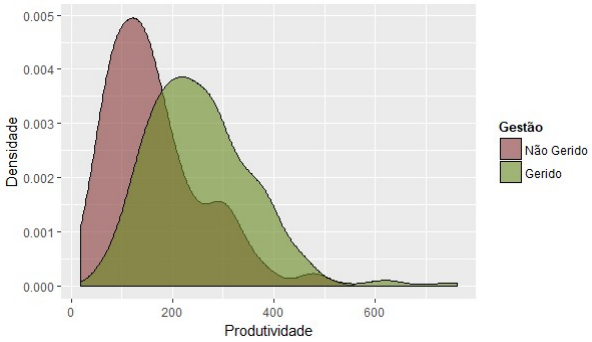
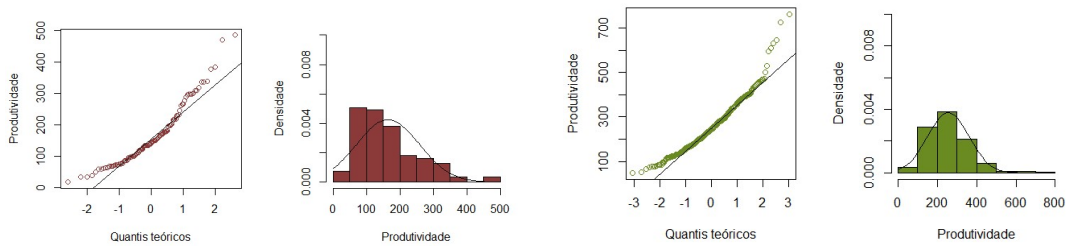


Figura 2: Gráfico relativo às curvas da distribuição de probabilidade estimada das parcelas geridas e não geridas.



(a) (b)

Figura 3: Gráficos *qqplot* e histogramas dos grupos de parcelas (a) sem evidências de gestão e (b) com evidências de gestão, relativamente à variável produtividade.

Relativamente à variável produtividade no grupo de parcelas Geridas, na figura 2, verifica-se que a curva da função densidade de probabilidade possui um pico, seguido de um decrescimento de ambos os lados, o que também é visível com base no histograma de frequências apresentado. Não obstante, contrariamente ao que sucede na curva da Distribuição Normal, a curva da função densidade estimada de probabilidade da produtividade neste grupo de parcelas não é simétrica, tendo por base as figuras 2 e 3a. Também o afastamento que se verifica entre os quantis teóricos e os quantis empíricos da distribuição Normal, visível no *qqplot*, torna plausível conjeturar a inexistência de Normalidade em relação à produtividade neste grupo de parcelas.

Relativamente à variável produtividade nas parcelas não geridas, a figura 2 permite intuir a existência de um pico na função densidade de probabilidade, seguido de um decrescimento de ambos os lados, sendo que este não é simétrico, o que contraria a existência de Normalidade nesta variável no grupo em estudo. Tal coaduna-se com a análise do *qqplot*, onde o afastamento existente entre os quantis teóricos da distribuição Normal e os quantis empíricos da amostra, corroboram com a conjetura de inexistência de Normalidade na variável produtividade no grupo de parcelas não geridas.

A análise dos resultados obtidos por aplicação dos testes de Normalidade considerados permite concluir a inexistência de Normalidade da variável produtividade relativamente aos dois grupos em estudo: parcelas Geridas e Não Geridas (tabela 1).

Tabela 1: Resultados do *p-value* obtidos por aplicação de testes de Normalidade aos dois grupos de parcelas - geridas e não geridas - relativamente à variável produtividade.

	Parcelas Geridas	**Parcelas Não Geridas**
Teste de Shapiro-Wilk	$1.206e-08$	$8.111e-05$
Teste de Anderson-Darling	$3.989e-05$	$3.235e-05$
Teste de KS com correção de Lilliefors	0.014	0.0005

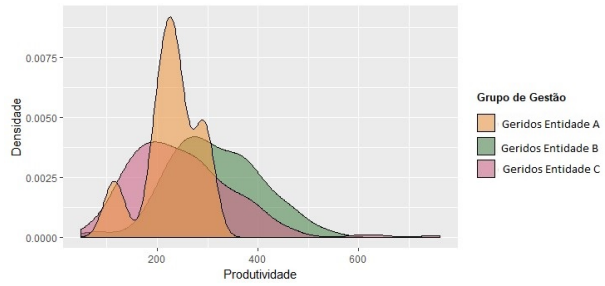
Homogeneidade dos grupos de parcelas geridas e não geridas

A implementação dos testes de Levene e de Fligner permitem a obtenção dos resultados sintetizados na tabela 2. Os valores do *p-value* superiores a 0.05, permitem intuir, a um nível de significância de 5%, a existência de homogeneidade das variâncias entre os grupos de parcelas com e sem evidencias da existência de gestão no que infere à variável produtividade.

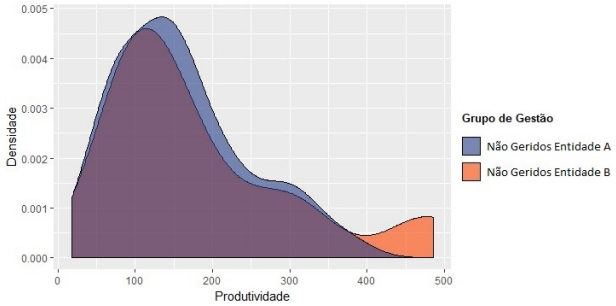
Tabela 2: Resultados obtidos por aplicação dos testes de homogeneidade das variâncias relativas à variável produtividade nos dois grupos de parcelas - geridas e não geridas.

	p-value	**Decisão** ($\alpha = 0.05$)
Teste de Levene	0.142	Não rejeitar homogeneidade das variâncias.
Teste de Fligner	0.096	Não rejeitar homogeneidade das variâncias.

D.3 Função densidade de probabilidade estimada da variável produtividade para os cinco grupos de gestão



(a)



(b)

Figura 4: Gráfico relativo à função densidade de probabilidade da variável produtividade relativa aos cinco grupos de parcelas formados.(a) Grupos de parcelas geridas. (b) Grupos de parcelas não geridas.

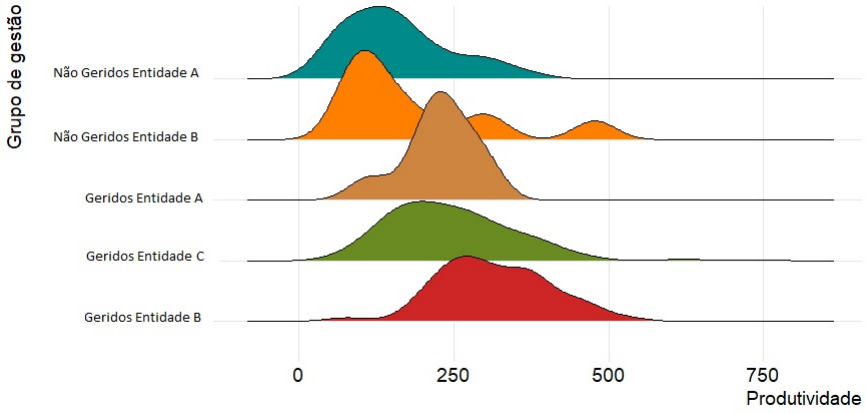


Figura 5: Estimativa da função densidade de probabilidade estimada da variável produtividade por grupo de gestão.

D.4 Função implementada no *software* R que aplica o método ANOVA a dois fatores não paramétrica

```
ANOVA2WayNonParametric<-function(Ordem, fa, fb, a, b, n, N, alpha){

  ###SQOFA
  p1_SQOFA<-aggregate(Ordem, by=list(Category=fa), FUN=sum)
  #print(p1_SQOFA)
  p2_SQOFA<-vector()
  for(i in 1:nrow(p1_SQOFA)) {
    p2_SQOFA[i]<-p1_SQOFA$x[i]^2
  }
  #print(p2_SQOFA)
  p3_SQOFA=sum(p2_SQOFA)
  SQOFa=(p3_SQOFA/(b*n))-((N*(N+1)^2)/4)
  sprintf("SQOFA= %s", SQOFa)
  #print(paste0("SQOFA= ", SQOFa))


  ###SQOFB
  p1_SQOFB<-aggregate(Ordem, by=list(Category=fb), FUN=sum)
  p2_SQOFB<-vector()
  for(i in 1:nrow(p1_SQOFB)) {
    p2_SQOFB[i]<-p1_SQOFB$x[i]^2
  }
  p3_SQOFB=sum(p2_SQOFB)
  SQOFb=(p3_SQOFB/(a*n))-((N*(N+1)^2)/4)
  sprintf("SQOFB= %s", SQOFb)
  #print(SQOFb)
  #print(p2)

  ###SQOAmostras
  p1_SQOAmostras<-aggregate(Ordem, by=list(Category=fa, fb), FUN=sum)
  p2_SQOAmostras<-vector()
  for(i in 1:nrow(p1_SQOAmostras)) {
    p2_SQOAmostras[i]<-p1_SQOAmostras$x[i]^2
  }
  p3_SQOAmostras=sum(p2_SQOAmostras)
  SQOAm=(p3_SQOAmostras/(n))-((N*(N+1)^2)/4)
  sprintf("SQOAm= %s", SQOAm)
  #print(SQOAm)
  #print(p2)

  ###SQO_axb
  SQO_axb=SQOAm-SQOFa-SQOFb
  sprintf("SQO_axb= %s", SQO_axb)

  ###QMOT
  QMOT<-(sum(Ordem^2)-(sum(Ordem)^2)/N)/(N-1)
  sprintf("QMOT= %s", QMOT)
  #print(QMOT)

  ##Calculo de H
  H_A<-round(SQOFa/QMOT,3)
  H_B<-round(SQOFb/QMOT,3)
  H_AxB<-round(SQO_axb/QMOT,3)

  ##Graus de liberdade
  gl_a<-a-1
  gl_b<-b-1
  gl_axb<-(a-1)*(b-1)
  gl_amostras<-(a*b)-1

  ##Obter p-values
  pval_Fa<-round(pchisq(H_A, df=gl_a, lower.tail=FALSE), digits=3)
  pval_Fb<-round(pchisq(H_B, df=gl_b, lower.tail=FALSE), 3)
  pval_AxB<-round(pchisq(H_AxB, df=gl_axb, lower.tail=FALSE),3)

  ##Escolher rejeicao/aceitacao da hipotese nula
  Decisao_A<-ifelse(H_A>=qchisq(1-alpha, gl_a), "Rejeitar", "Nao rejeitar")
  Decisao_B<-ifelse(H_B>=qchisq(1-alpha, gl_b), "Rejeitar", "Nao rejeitar")
  Decisao_AxB<-ifelse(H_AxB>=qchisq(1-alpha, gl_axb), "Rejeitar", "Nao rejeitar")

  FontVariacao<-c("Fator_A", "Fator_B", "Interacao", "Amostras")
  SO<-c(SQOFa, SQOFb, SQO_axb, SQOAm)
  GL<-c(gl_a, gl_b, gl_axb, gl_amostras)
  H<-c(H_A, H_B, H_AxB, "")
  pval<-c(pval_Fa, pval_Fb, pval_AxB, "")
  Decisao<-c(Decisao_A, Decisao_B, Decisao_AxB, "")
  result<-data.frame(FontVariacao, SO, GL, H,pval, Decisao)


  print(result)
}
```

D.5 Testes de hipóteses aplicados sob as idades das parcelas.

Homogeneidade das variâncias das idades entre grupos.

Tabela 3: Resultados da aplicação dos testes de homogeneidade das variâncias sob as idades das parcelas nos grupos.

	Teste de Levene	**Teste de Fligner**	**Decisão** ($\alpha = 0.05$)
Entre as idades das parcelas Geridas e Não Geridas	9.945×10^{-8}	5.784×10^{-8}	Rejeitar a existência de homogeneidade de variância das idades
Entre as idades das parcelas $NGeridas_B$ e $NGeridas_A$	0.783	0.800	Não rejeitar existência de homogeneidade de variância das idades
Entre as idades das parcelas $Geridas_B$ e $Geridas_A$	0.908	0.936	Não rejeitar existência de homogeneidade de variância das idades
Entre as idades das parcelas $Geridas_B$ e $Geridas_C$	0.011	0.0014	Rejeitar a existência de homogeneidade de variância das idades
Entre as idades das parcelas $Geridas_C$ e $Geridas_A$	0.330	0.223	Não rejeitar existência de homogeneidade de variância das idades

Resultados da aplicação do teste de Wilcoxon-Mann-Whitney (unilateral) às idades das parcelas Geridas pelas 3 entidades A, B e C.

A aplicação do teste não paramétrico justifica-se tendo por base a falta de Normalidade inerente aos grupos, relativamente à variável idade das parcelas.

Tabela 4: Resultados da aplicação do teste de Wilcoxon-Mann-Whitney sob a idade das parcelas das parcelas geridas pertencentes às 3 entidades A, B e C, onde θ representa a mediana das idades das parcelas.

Hipóteses do teste	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0: \ \theta_{Geridos_B} = \theta_{Geridos_A}$ vs $H_1: \ \theta_{Geridos_B} < \theta_{Geridos_A}$	0.713	Não rejeitar H_0
$H_0: \ \theta_{Geridos_B} = \theta_{Geridos_C}$ vs $H_1: \ \theta_{Geridos_B} < \theta_{Geridos_C}$	1.611×10^{-7}	Rejeitar H_0
$H_0: \ \theta_{Geridos_A} = \theta_{Geridos_C}$ vs $H_1: \ \theta_{Geridos_A} < \theta_{Geridos_C}$	0.002	Rejeitar H_0

D.6 Resultados das análises relativas à variável seleção de varas.

Contabilização de parcelas por grupo de gestão

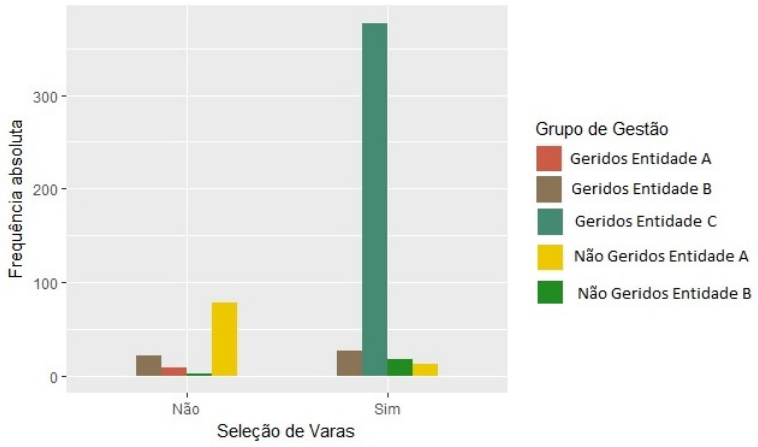


Figura 6: Gráfico de barras relativo à frequência absoluta de parcelas, em função da aplicação ou não de seleção de varas, por cada grupo de gestão.

Tabela 5: Contabilização do número de parcelas com e sem seleção de varas por grupo de gestão em frequência absoluta.

	Seleção de Varas	
	Com seleção	**Sem Seleção**
Geridos Entidade A	8	0
Geridos Entidade B	21	27
Geridos Entidade C	0	377
Não Geridos Entidade A	78	13
Não Geridos Entidade B	2	17

Comparação de medianas entre os grupos de parcelas geridas e não geridas, com e sem seleção de varas

Tabela 6: Resultados da aplicação do teste de Wilcoxon-Mann-Whitney sob os dados geridos e não geridos, com e sem seleção de varas.

Hipóteses do teste	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0: \ \theta_{Geridos_{ComSelecãoVaras}} = \theta_{NGeridos_{ComSelecãoVaras}}$ vs $H_1: \ \theta_{Geridos_{ComSelecãoVaras}} > \theta_{NGeridos_{ComSelecãoVaras}}$	0.0056	Rejeitar H_0
$H_0: \ \theta_{Geridos_{ComSelecãoVaras}} = \theta_{NGeridos_{SemSelecãoVaras}}$ vs $H_1: \ \theta_{Geridos_{ComSelecãoVaras}} > \theta_{NGeridos_{SemSelecãoVaras}}$	$< 2.2e-16$	Não rejeitar H_0
$H_0: \ \theta_{Geridos_{ComSelecãoVaras}} = \theta_{Geridos_{SemSelecãoVaras}}$ vs $H_1: \ \theta_{Geridos_{ComSelecãoVaras}} < \theta_{Geridos_{SemSelecãoVaras}}$	0.024	Rejeitar H_0
$H_0: \ \theta_{NGeridos_{ComSelecãoVaras}} = \theta_{Geridos_{SemSelecãoVaras}}$ vs $H_1: \ \theta_{NGeridos_{ComSelecãoVaras}} < \theta_{Geridos_{SemSelecãoVaras}}$	0.0013	Rejeitar H_0
$H_0: \ \theta_{NGeridos_{ComSelecãoVaras}} = \theta_{NGeridos_{SemSelecãoVaras}}$ vs $H_1: \ \theta_{NGeridos_{ComSelecãoVaras}} > \theta_{NGeridos_{SemSelecãoVaras}}$	$2.63e-9$	Rejeitar H_0

D.7 Resultados das análises relativas à variável preparação do terreno.

Contabilização de tipo de preparação do terreno por grupo de gestão

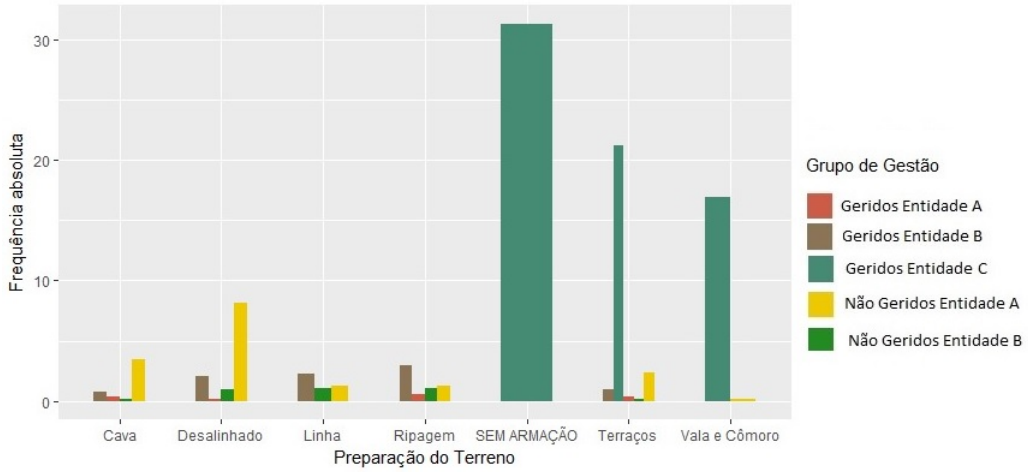


Figura 7: Gráfico de barras relativo à frequência absoluta de parcelas, em função do tipo de preparação do terreno aplicado por cada grupo de gestão.

Tabela 7: Contabilização do número de parcelas em função do grupo de gestão e do tipo de preparação do terreno aplicado em frequência absoluta.

	Preparação do terreno						
	Cava	**Desalinhado**	**Linha**	**Ripagem**	**Sem armação**	**Terraços**	**Vala e Comoro**
$Geridos_A$	2	1	0	3	0	2	0
$Geridos_B$	4	11	12	16	0	5	0
$Geridos_C$	0	0	0	0	170	115	92
$NGeridos_A$	19	44	7	7	0	13	1
$NGeridos_B$	1	5	6	6	0	1	0

Comparação da mediana da produtividade entre parcelas geridas e não geridas, em função do tipo de preparação do terreno aplicado. θ designa a mediana da produtividade das parcelas.

Tabela 8: Resultados da aplicação do teste de Wilcoxon-Mann-Whitney sob os dados geridos e não geridos, em função do tipo de preparação do terreno aplicado.

Hipótese do teste	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0: \ \theta_{Cava_{NGeridos}} = \theta_{Cava_{Geridos}}$ vs $H_1: \ \theta_{Cava_{NGeridos}} < \theta_{Cava_{Geridos}}$	0.0097	Rejeitar H_0
$H_0: \ \theta_{Desalinhado_{NGeridos}} = \theta_{Desalinhado_{Geridos}}$ vs $H_1: \ \theta_{Desalinhado_{NGeridos}} < \theta_{Desalinhado_{Geridos}}$	$3.006e-06$	Rejeitar H_0
$H_0: \ \theta_{Linha_{NGeridos}} = \theta_{Linha_{Geridos}}$ vs $H_1: \ \theta_{Linha_{NGeridos}} < \theta_{Linha_{Geridos}}$	0.001	Rejeitar H_0
$H_0: \ \theta_{Ripagem_{NGeridos}} = \theta_{Ripagem_{Geridos}}$ vs $H_1: \ \theta_{Ripagem_{NGeridos}} < \theta_{Ripagem_{Geridos}}$	0.009	Rejeitar H_0
$H_0: \ \theta_{Sem\ armacão_{Geridos}} = \theta_{Sem\ armacão_{NGeridos}}$ vs $H_1: \ \theta_{Sem\ armacão_{Geridos}} < \theta_{Sem\ armacão_{NGeridos}}$	*	
$H_0: \ \theta_{Terracos_{Geridos}} = \theta_{Terracos_{NGeridos}}$ vs $H_1: \ \theta_{Terracos_{Geridos}} < \theta_{Terracos_{NGeridos}}$	$1.068e-05$	Rejeitar H_0
$H_0: \ \theta_{Vala_{Geridos}} = \theta_{Vala_{NGeridos}}$ vs $H_1: \ \theta_{Vala_{Geridos}} < \theta_{Vala_{NGeridos}}$	*	

*Não é possível comparação por falta de representatividade das parcelas não geridas.

D.8 Resultados das análises relativas à variável controlo da vegetação

Contabilização do número de parcelas com e sem controlo de gestão, por grupo de gestão formado

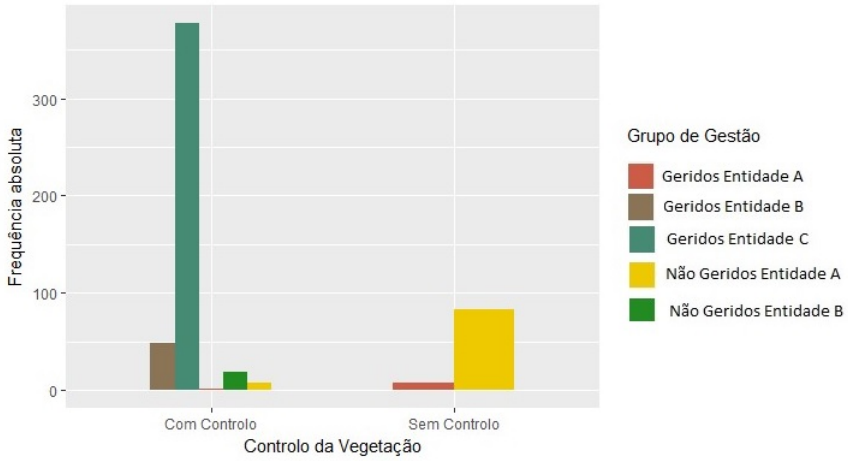


Figura 8: Gráfico de barras relativo à frequência absoluta de parcelas, em função da existência de controlo da vegetação, por cada grupo de gestão.

Tabela 9: Contabilização do número de parcelas com e sem controlo da vegetação por grupo de gestão em frequência absoluta.

	Controlo da vegetação	
	Com controlo	**Sem controlo**
$Geridos_A$	1	7
$Geridos_B$	48	0
$Geridos_C$	377	0
$NGeridos_A$	8	83
$NGeridos_B$	19	0

Comparação da mediana da produtividade entre parcelas geridas e não geridas em função da existência de controlo da vegetação

Tabela 10: Resultados da aplicação do teste de Wilcoxon-Mann-Whitney sob os dados geridos e não geridos, com e sem controlo da vegetação.

Hipóteses do teste	***p-value***	**Decisão** ($\alpha = 0.05$)
$H_0: \ \theta_{Geridos_{ComControlo}} = \theta_{NGeridos_{ComControlo}}$ vs $H_1: \ \theta_{Geridos_{ComControlo}} > \theta_{NGeridos_{ComControlo}}$	$1.85e-5$	Rejeitar H_0
$H_0: \ \theta_{Geridos_{ComControlo}} = \theta_{NGeridos_{SemControlo}}$ vs $H_1: \ \theta_{Geridos_{ComControlo}} > \theta_{NGeridos_{SemControlo}}$	$4.36e-15$	Rejeitar H_0
$H_0: \ \theta_{Geridos_{ComControlo}} = \theta_{Geridos_{SemControlo}}$ vs $H_1: \ \theta_{Geridos_{ComControlo}} < \theta_{Geridos_{SemControlo}}$	0.740	Não rejeitar H_0
$H_0: \ \theta_{NGeridos_{ComControlo}} = \theta_{Geridos_{SemControlo}}$ vs $H_1: \ \theta_{NGeridos_{ComControlo}} < \theta_{Geridos_{SemControlo}}$	0.037	Rejeitar H_0
$H_0: \ \theta_{NGeridos_{ComControlo}} = \theta_{NGeridos_{SemControlo}}$ vs $H_1: \ \theta_{NGeridos_{ComControlo}} < \theta_{NGeridos_{SemControlo}}$	0.644	Não rejeitar H_0
$H_0: \ \theta_{Geridos_{SemControlo}} = \theta_{NGeridos_{SemControlo}}$ vs $H_1: \ \theta_{Geridos_{SemControlo}} > \theta_{NGeridos_{SemControlo}}$	0.014	Rejeitar H_0

D.9 Resultados obtidos por aplicação de árvores de regressão com re-amostragem e validação cruzada, mas sem procedimentos de *tuning*.

Código usado na construção da árvore do modelo 1

- Construção de árvore regressão do modelo 1 sem poda.

```
library(rpart)

# Uso do metodo anova
arvore1 <- rpart(Produtividade ~ FR_n_vivas + Idade_validada +as.factor(Rotacao) + as.factor(RP) +
as.factor(entd), method="anova", data=dados.tr)

#Apresentar resultados
print(arvore1)
summary(arvore1) #resultados mais detalhados
```

- Código usado na construção da árvore de regressão do modelo 1 com aplicação de metodologias de poda.

```
#Aplicar CV ao modelo arvore1
printcp(arvore1) #para avaliar valores numericos
plotcp(arvore1) # visualizar graficamente os resultados

###Determinar valor otimo para podar a arvore
bestcp <- arvore1$cptable[which.min(arvore1$cptable[,"xerror"]),"CP"]

# Com o valor assim determinado para cp, obtem-se uma nova arvore podada, fazendo:
tree.pruned <- prune(arvore1, cp = bestcp)

#Representar arvore podada
plot(arv.produtividade, uniform=TRUE,
     main="Regression Tree for Mileage ")
text(arv.produtividade, use.n=TRUE, all=TRUE, cex=.8)

#Outra forma de representar
prp(tree.pruned, faclen = 0, cex = 0.8, extra = 1)

summary(tree.pruned) #ver valores detalhados da arvore podada

pred<-predict(tree.pruned, data=dados.te) #obter previsoes
RMSE(pred, dados.te$Produtividade) #calculo de RMSE
```

Código usado na construção da árvore de regressão do modelo 2

- Código usado na construção da árvore de regressão do modelo 2, aplicando metodologias de *tunning*, sem aplicação de poda.

```
hyper_grid <- expand.grid(minsplit = seq(0, 30, 1), maxdepth = seq(0, 50, 1))

models <- list()

for (i in 1:nrow(hyper_grid)) {
  # Obter minsplit, maxdepth da linha i
  minsplit <- hyper_grid$minsplit[i]
  maxdepth <- hyper_grid$maxdepth[i]

  # Treino do modelo e guardar os resultados na lista criada
  models[[i]] <- rpart(
    formula = Produtividade ~ FR_n_vivas + Idade_validada + as.factor(Rotacao) +
      as.factor(RP) + as.factor(entd), data=dados.tr,
    method  = "anova",
    control = list(minsplit = minsplit, maxdepth = maxdepth))
}

# Funcao para obter o melhor valor de CP
get_cp <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  cp <- x$cptable[min, "CP"]
}

```

```
# Funcao para obter o minimo valor do erro
get_min_error <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  xerror <- x$cptable[min, "xerror"]
}

#Filtrar os valores do hyper_grid com base nos valores otimos obtidos para cp e para o erro minimo
hyper_grid %>%
  mutate(
    cp    = purrr::map_dbl(models, get_cp),
    error = purrr::map_dbl(models, get_min_error)
  ) %>%
  arrange(error) %>%
  top_n(-5, wt = error)

###Resultados:
# minsplit maxdepth         cp     error
#1       27       39 0.02099329 0.6889116
#2        5       14 0.01880731 0.6928393
#3       16       23 0.02099329 0.6940579
#4       14       49 0.02099329 0.6954532
#5       13       43 0.02099329 0.6969483


#Construcao da arvore otima cm base nos valores otimos obtidos para controlo
optimal_tree <- rpart(formula = Produtividade ~ FR_n_vivas + Idade_validada + as.factor(Rotacao) + as.factor
     (RP) + as.factor(entd), data    = dados.tr, method  = "anova", control = list(minsplit = 27, maxdepth =
      39, cp = 0.02))

####Representar a arvore otima:
plot(optimal_tree, uniform=TRUE)
text(optimal_tree, use.n=TRUE, all=TRUE, cex=.8)
```

- Código usado na construção da árvore de regressão do modelo 2, fazendo uso de técnicas de *tunnig* e poda da árvore.

```

###Determinar valor otimo para podar a arvore
bestcp2 <- optimal_tree$cptable[which.min(optimal_tree$cptable[,"xerror"]),"CP"]

# Obter arvore podada
tree.pruned_arv2 <- prune(optimal_tree, cp = bestcp2)

#Validacao e avaliacao do modelo da arvore 2 podada
pred2 <- predict(tree.pruned_arv2, newdata = dados.te)
RMSE(pred = pred2, obs = dados.te$Produtividade)

#representar a arvore 2 podada
plot(tree.pruned_arv2)
text(tree.pruned_arv2, cex = 0.8, use.n = TRUE, xpd = TRUE)
```